The Person Response Curve: Fit of Individuals to Item Characteristic Curve Models

Tom E. Trabin
and
David J. Weiss

Research Report 79-7
December 1979

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

This research was supported by funds from the Navy Personnel Research and Development Center, and Office of Naval Research, and monitored by the Office of Naval Research.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER 9 14 RR
Research Rep~~ort~~ 79-7

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle) 6
The Person Response Curve: The Fit of Individuals to Item Characteristic Curve Models.

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) 10
Tom E. Trabin ~~and~~ David J. Weiss

8. CONTRACT OR GRANT NUMBER(s) 15
N00014-76-C-0243

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P.E.:61153N PROJ.:RR042-04
T.A.:RR042-04-01
W.U.:NR150-382

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE 11
Dec~~ember~~ 1979

13. NUMBER OF PAGES
36

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 45

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.
16 RR04204

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
17 RR0420401

18. SUPPLEMENTARY NOTES
This research was supported by funds from the Navy Personnel Research and Development Center, and the Office of Naval Research, and monitored by the Office of Naval Research.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

intra-individual dimensionality
three-parameter logistic
latent trait test theory
item characteristic curve theory
carelessness
guessing
testwiseness
ability testing
deviant responses
achievement testing
adaptive testing
tailored testing
person-fit

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study investigated a method of determining the fit of individuals to item characteristic curve (ICC) models using the person response curve (PRC). The construction of observed PRCs is based on an individual's proportion correct on test item subsets (strata) that differ systematically in difficulty level. A method is proposed for identifying irregularities in an observed PRC by comparing it with the expected PRC predicted by the three-parameter ICC logistic model for that individual's ability level. Diagnostic

406 024

1B

potential of the PRC is discussed in terms of the degree and type of deviations of the observed PRC from the expected PRC predicted by the model.

Observed PRCs were constructed for 151 college students using vocabulary test data on 216 items of wide difficulty range. Data on students' test-taking motivation, test-taking anxiety, and perceived test difficulty were also obtained. PRCs for the students were found to be reliable and to have shapes that were primarily a function of ability level. Three-parameter logistic (ICC) model expected PRCs served as good predictors of observed PRCs for over 90% of the group. As anticipated from this general overall fit of the observed data to the ICC model, there were no significant correlations between degree of non-fit and test-taking motivation, test-taking anxiety, or perceived test difficulty. Using split-pool observed PRCs, a few students were identified who deviated significantly from the expected PRC.

The results of this study suggested that three-parameter logistic expected PRCs for given ability levels were good predictors of test response profiles for the students in this sample. Significant non-fit between observed and expected PRCs would suggest the interaction of additional dimensions in the testing situation for a given individual. Recommendations are made for further research on person response curves.

CONTENTS

Acknowledgments

The assistance of Isaac Bejar, Austin T. Church, J. Stephen Prestwood, and James B. Sympson in various aspects of the design of this study and the analysis of these data is gratefully acknowledged.

Technical Editor: Barbara Leslie Camm

The Person Response Curve: Fit of Individuals to Item Characteristic Curve Models

The development of group ability tests more than 50 years ago has enabled the comparison of the total test score of an individual with the scores of a population norm group, thus allowing for more meaningful interpretation of ability estimates than can be done with the use of simple number-correct scores. For example, the statement "On the XYZ aptitude test John scored at the 73rd percentile of college students" gives more information than the statement,"On the ABC ability test Mary correctly answered 64 questions out of 90, whereas Sam correctly answered 33 questions." Both examples have in common the report of a person's test performance on a specific dimension given in terms of an overall test score; but this single summary score, while more parsimonious than a description of a testee's entire response pattern, may not reveal the operation of other factors on test-taking behavior, such as guessing, anxiety, cultural bias, or lack of motivation. Thus, total scores on a test do not indicate whether that test is inappropriate for a certain individual or group of individuals.

The emergence of modern test theory, based on the item characteristic curve (ICC; Hambleton & Cook, 1977; Lord & Novick, 1968), brings with it the promise of better tests conveying more accurate information about testee ability levels. This is partially accomplished by use of ability estimation procedures that take into account the testee's total response pattern in estimating ability levels (Bejar & Weiss, 1979; Kingsbury & Weiss, 1979a). These scoring methods can provide individualized error bands around the testee's ability level estimates, which indicate the precision of those ability estimates (e.g., Kingsbury & Weiss, 1979b). Thus, in addition to providing methods designed to permit more adequate test construction by the use of test information curves (Hambleton & Cook, 1977), ICC theory permits utilizing more of a testee's response pattern in order to provide individualized estimates of precision for ability estimates. In addition, ICC theory also allows for the development of powerful methods of adaptive testing for the solution of many practical measurement problems (e.g., Brown & Weiss, 1977; Kingsbury & Weiss, 1979b; McBride & Weiss, 1976; Vale & Weiss, 1977; Weiss, 1973, 1975).

In contrast to classical test theory, ICC theory makes strong assumptions in order to achieve its objectives. The major operational forms of ICC theory assume 1) local independence, 2) unidimensionality, and 3) a specified shape for the item characteristic curve. Although local independence cannot be directly demonstrated, data supporting the unidimensionality assumption in a variety of settings (e.g., Bejar, Weiss, & Kingsbury, 1977; Church, Pine, & Weiss, 1978; Martin, Pine, & Weiss, 1978; McBride & Weiss, 1974; Reckase, 1978) lend indirect support to the assumption of local independence. Lord (1968) has presented data showing that the assumption of a normal ogive ICC is tenable and, given the minor differences between a logistic ogive and a normal ogive, has indirectly supported the use of the logistic item response function in ICC theory.

There has been very little research, however, to demonstrate that *individuals* behave in accordance with the ICC model, although a growing concern has been exhibited in the testing literature for the development of methods to extract more information from test response data than simply a total score. Use of ICC models with individuals must rest on a demonstration that the test responses of individuals are in accordance with the testing model hypothesized. If this can be demonstrated for most individuals on a number of tests, ICC models can be used with confidence to their full power. On the other hand, if a majority of individuals respond in ways contrary to ICC theory, the utility of the theory for individual measurement can be seriously questioned.

A major advantage of the assumptions of ICC theory for individual measurement is that the question of individuals' fit or non-fit to the model can be investigated on an individual basis. The practical implications of identifying non-fitting persons were realized by Educational Testing Service in their study of methods to identify response patterns of the type of student who "may be so atypical and unlike other students that his [or her] aptitude test score fails to be a completely appropriate measure of his [or her] relative ability" (Levine & Rubin, 1976). Examples of such students are low-ability examinees who copy answers to several difficult items from a much more able neighbor and very high-ability examinees fluent in another language but not yet fluent in English, who misunderstand the wording of several relatively easy questions. Levine and Rubin recommended the development of indices to identify such test item response patterns as a "rich and fertile area for future research."

The appropriateness of a certain test or certain items for specific individuals has also been an important concern for test developers working with the one-parameter logistic ICC (Rasch) model. Wright and his associates (1977; Mead, 1979; Wright & Stone, 1979; Wainer & Wright, in prep.) have proposed identification of such factors as guessing, carelessness, and bias,using the Rasch model. According to Lumsden (1977), a bright but careless student may have the same overall ability score as a careful and consistent average student, but there are differential instructional implications for teaching these two types of students or differential counseling implications if the two students are seeking vocational counseling.

Thus, the question of fit or non-fit of individuals to ICC testing models has important practical and theoretical importance. Fit of individuals must be demonstrated in order to realize the full potential of the model for practical use. At the same time, the development of reliable and valid methods for quantifying and identifying aberrant response patterns would provide a potentially useful source of additional information on test-taking behavior of individuals.

Related Research

The question of fit of individuals to the ICC models can be conceptualized as investigating the variability of a single individual in a single testing situation. Wright (1977), in suggesting that to postulate and to study such a phenomonon would be to "wreak havoc with the logic and practice of measurement," exemplifies an attitude which may, in part, account for the meager literature on the topic. It is more likely, however, that the development of sufficiently refined measurement techniques to handle such a difficult problem has not occurred until very recently. The development of computerized testing together with the development of latent trait test theory was necessary to bring about the possibility of measuring individual variablity with a single test.

Most of the existing research consists of tentative theoretical approaches with closing exhortations for further study. The approaches to this problem differ widely in theoretical orientation and in terminology used. Mosier (1942) first referred to individual variability in mental test theory from a psychophysical orientation; and Levine and Rubin (1976) referred to aberrance indices from the view of signal detection theory. Lumsden (1977) used the Thurstoneian approach of categorical judgment to propose the idea of person reliability. Weiss (1973) used data from adaptive testing to develop consistency scores, and Vale and Weiss (1975) further developed the earlier idea of consistency scores into an empirical study of subject characteristic curves. Wright (1977) used the one-parameter logistic model to propose the idea of item residuals and to refute the notion of what he called person sensitivity in testing. Clearly, the idea is still new, hazily formulated on a theoretical level, with very scarce evidence of any empirical studies.

Mosier's psychophysical approach. The first reference in the testing literature to an individual's variability within a single ability testing situation was in a two-part study by Mosier (1940, 1942). The emphasis in this study was on the fundamental relationships between the field of mental test theory and the methods of measuring psychophysical processes. This comparison included relating the constant method of psychophysical measurement with scoring by the number-correct method in mental testing. Mosier asserted that a composite score is an imperfect representation of an individual's test score and depends on the individual's variability, just as an individual's threshold in psychophysics depends on the ambiguity of the stimulus; as a stimulus is variable with respect to a group of judges, so an individual is variable with respect to a group of items.

Mosier likened the ambiguity (discriminal dispersion) of a stimulus in psychophysics to individual variability in mental test performance. He postulated the distribution of the proportion of correct answers for one individual across items of differing difficulty as the integral of the normal probability curve and the variability of that individual as the standard deviation of the probability function whose integral is the proportion of correct answers as a function of difficulty. Mosier applied the constant process of psychophysics to a set of test data (of unspecified characteristics) and estimated the difficulty of median error for individuals (ability level) and its dispersion. He found odd-even reliability of ability level estimated by this method to be .88. The reliability of the person variability index was .55, a value significantly different from zero. It was perhaps this apparent low reliability estimate which was responsible for a complete lack of research on person variability for the next 30 years.

Weiss's stradaptive "trace line". The idea of person variability within one test was independently developed by Weiss (1973) as a by-product of computerized adaptive testing. In the design of the stratified-adaptive (stradaptive) test, he ordered ability test items by difficulty levels into strata. In examining testee performance on stradaptive tests, Weiss noted that individuals who correctly answered items of the same average difficulty level differed in terms of the proportion of items they answered correctly at different difficulty levels.

To examine differences in individual variability, Weiss proposed the concept of a "trace line" for a testee's item responses, with items divided into strata of increasing difficulty on the x-axis and proportion correct for an individual on each stratum on the y-axis, duplicating the suggestion of Mosier 30 years earlier. Weiss hypothesized as did Mosier, that proportion correct would decrease as stratum difficulty increased. Also echoing Mosier, he proposed that the steepness of the slope be interpreted as an index of the consistency of an individual's item responses and the capability of the item pool to discriminate an individual's ability level. The point of inflection of the curve, where 50% of the items were answered correctly (for free-response items) was proposed as an indicator of the difficulty of the item pool for an individual or the position of that individual on the trait continuum. To operationalize the concept of person variability, or what Weiss called "consistency," he suggested calculating several indices, including the standard deviation of item difficulties answered correctly and the standard deviation of item difficulties encountered.

Vale and Weiss (1975) empirically studied some aspects of individual "consistency" as part of a larger study of computer-administered adaptive testing. Included in this study was a test of the hypothesis that more consistent individuals--those with smaller errors of measurement in Mosier's (1940, 1942) formulation--would have more stable ability estimates. The five operationalizations of consistency originally proposed by Weiss (1973) were studied as moderators in the prediction of test-retest reliability of ability estimates. The standard deviation of item difficulties encountered significantly moderated the stability of ability estimates in the expected direction as, to a lesser extent, did the standard deviation of item difficulties answered correctly.

In addition, Vale and Weiss (1975) studied the test-retest reliability of the "trace line" plots for individuals and introduced the new term "subject characteristic curve" for these trace lines. They used canonical redundancy analysis (Weiss, 1972) on the proportion-correct-within-strata data (i.e., the subject characteristic curves) in a retest situation. The results indicated a high degree of predictability of subject characteristic curves on one test from the test scores on the other; redundancies indicated from 47% to 67% common variance across the two testing times. These results indicated a good degree of stability in the proportion of correct responses within the strata of the stradaptive test as indexed by the subject characteristic curves.

Lumsden's subject characteristic curve. The subject characteristic curve was again independently proposed by Lumsden (1977, 1978) as a derivation from Thurstone's law of categorical judgment. Lumsden proposed an attribute-based model of test performance in which a person's ability fluctuates in trends (long-term developmental changes), swells (short-term mood swings), and tremors (moment-to-moment shifts). He assumed tremors to be rapid, random, and normally distributed shifts of ability occurring from moment to moment within a single test situation: The discriminal dispersion of item difficulties stays at zero, and it is only person ability that fluctuates. If the momentary location of a person's ability level is higher than the point location of the item's difficulty, the person will answer an item correctly. If ability is lower at any moment than the item difficulty location, the person will answer that item incorrectly. Lumsden then extended the idea to the plot, for a single person, of item responses at different difficulty levels, which he called the "person characteristic curve." He suggested that the person characteristic curve is "perfectly analogous to the item characteristic curve." Lumsden's basic assumptions, however, are different from the ICC theory assumptions underlying item characteristic curves; an ICC assumes that ability level is constant, not fluctuating, but that the response to a given test item includes a random error component causing observed item responses to fluctuate around true ability level.

Levine and Rubin's aberrancy indices. Other approaches to the study of intra-individual variability within a test have concentrated on the use of intra-individual variability for test validation rather than on individual ability assessment. Levine and Rubin (1976) and Levine (1979) initiated several studies concerned with individuals or groups of individuals for whom a given test might be invalid and/or inappropriate. Among the populations of concern were those who obtain higher scores because of cheating and those who obtain lower scores because of lack of proficiency in English. Levine and Rubin developed several types of "aberrance indices" to determine at greater than chance level, without reference to demographic data, examinees for whom a given test would be inappropriate.

Their basic assumption was that an aberrant examinee's response pattern to items of varying difficulty should have a low marginal probability, since it is unlikely that a high-ability examinee would incorrectly answer an easy item or a low-ability examinee correctly answer a difficult item. Marginal probability was operationally defined as the average of the conditional probabilities of a correct response on each item of difficulty level b for an individual of ability level θ. If n = the number of items, there are 2^n marginal probabilities. These were ranked, with all probabilities below an arbitrary cutoff point considered to represent aberrant response patterns.

Using a monte carlo simulation with 3,000 hypothetical examinees, 200 of whom were aberrant responders, Levine and Rubin (1976) conducted several studies at different cutoff points on the marginal probabilities to determine if aberrant examinees could be identified at a rate significantly greater than chance. Receiver operator curves (ROC) from signal detection theory were used to evaluate the performance of their experimental methods of identifying aberrance. The best method identified 50% of the spuriously low and 80% of the spuriously high examinees, while only mistaking 10% of the normal examinees as aberrant.

When compared to the chance level predictions of only 10% of spuriously high or low examinees identified while mistaking 10% of the normal examinees, this study seemed to have yielded results that merit further study. However, a closer look reveals the impracticality of Levine and Rubin's best method. Even if the aberrance indices identified 80% of the aberrant examinees (160 out of 200) and only misclassified 10% of the non-aberrant examinees (280 out of 2,800), this would still result in eliminating as invalid the test results of 280 non-aberrant examinees. Levine and Rubin seemed to completely ignore this problem in their paper.

Wright's residual analysis. Wright's (1977; Wright & Mead, 1977; Wright & Stone, 1979) concern with intra-individual variability in a single situation focuses on the interaction of a person with specific test items. Wright has developed methods for identifying items which may be invalid for a certain person or group of persons and which can then be excluded from consideration when calculating ability estimates from those items. Wright (1977; Wright & Stone, 1979, pp. 165-180) cited tendencies such as guessing, cheating, "sleeping" (getting bored with a test and answering later items in a more haphazard fashion), "fumbling" (e.g., answering earlier items with difficulty because of confusion with test format), and cultural bias. Wright's method (Wright & Stone, 1979; Mead, in prep.) utilizes standardized squares of the residuals between an item's difficulty level and a person's ability level after fitting the one-parameter logistic model to the test data. If these residuals indicate a significantly low probability of responding correctly or incorrectly and the person responded in that way, the tester then has reason to suspect that the item or item set may be invalid for that particular person.

This approach is consistent with Wright's use of the one-parameter Rasch model, which recognizes only a difficulty level of items but not a discrimination parameter or a guessing parameter. Following the assumptions of this model, Wright maintained that the probability of success on more difficult items should always be less than on easier items no matter who attempts the items, so the test developer must prevent variation in item discrimination sufficient to produce item characteristic curves that cross. Also, following this logic, a higher ability person should have a better chance for success no matter what the difficulty of the item attempted, so the test developer must prevent variation in person sensitivity; the result is that person characteristic curves must not cross each other. Wright claimed that the practical problem of variation in item discrimination and person sensitivity can be treated through supervision rather than estimation, using residuals and deleting inappropriate items from a person's responses without interfering with estimates of a person's ability. Wright's method seems to oversimplify response data by ignoring the effects of item discrimination and guessing, as well as precluding the possibility of more subtle diagnoses of added dimensions acting as moderator variables in the testing situation.

Summary and Objectives

The limited literature on person variability within a test thus seems to have three major trends: 1) the direct analysis of person variability as originally suggested by Mosier, later called the testee's trace line by Weiss and subject characteristic curve by Vale and Weiss and the person characteristic curve by Lumsden (1977); 2) designation of highly variable persons as aberrant by Levine and Rubin; and 3) the elimination of aberrant person-item interactions by Wright. Careful analysis of these three approaches indicates that the first approach (that of the person characteristic curve) is the most general of the three, subsuming the other two as special cases: If the entire pattern of a testee's responses is studied as a function of difficulty level of the items, the identification of aberrant response patterns or person-item restrictions follows directly. In addition, postulating a person characteristic curve in conjunction with ICC theory provides a means of testing for single individuals, whether their response patterns fit the theory regardless of the number of parameters assumed.

The purpose of this study was to further explore the Mosier-Weiss-Lumsden idea of the person characteristic curve, to determine its utility as a means of describing testee response variability, and to study the fit of individuals to the ICC model. To emphasize that the curve is derived from the responses of an individual to a set of test items, it was renamed the "person response curve." The focus of this research is on the investigation of the reliability and other psychometric characteristics of the person response curve.

The Person Response Curve

Observed Person Response Curves

Figure 1 is a plot of person response curves (PRCs) for each of three hypothetical testees. To obtain these plots, a number of items of different difficulty levels are administered to a testee. For each difficulty level, the proportion of items answered correctly is plotted as a function of difficulty level. The resulting PRC is representative of one person's performance on one test.

Figure 1
Observed Person Response Curves for Three Hypothetical Persons with the Same Ability Level ($\hat{\theta}$=0.0)

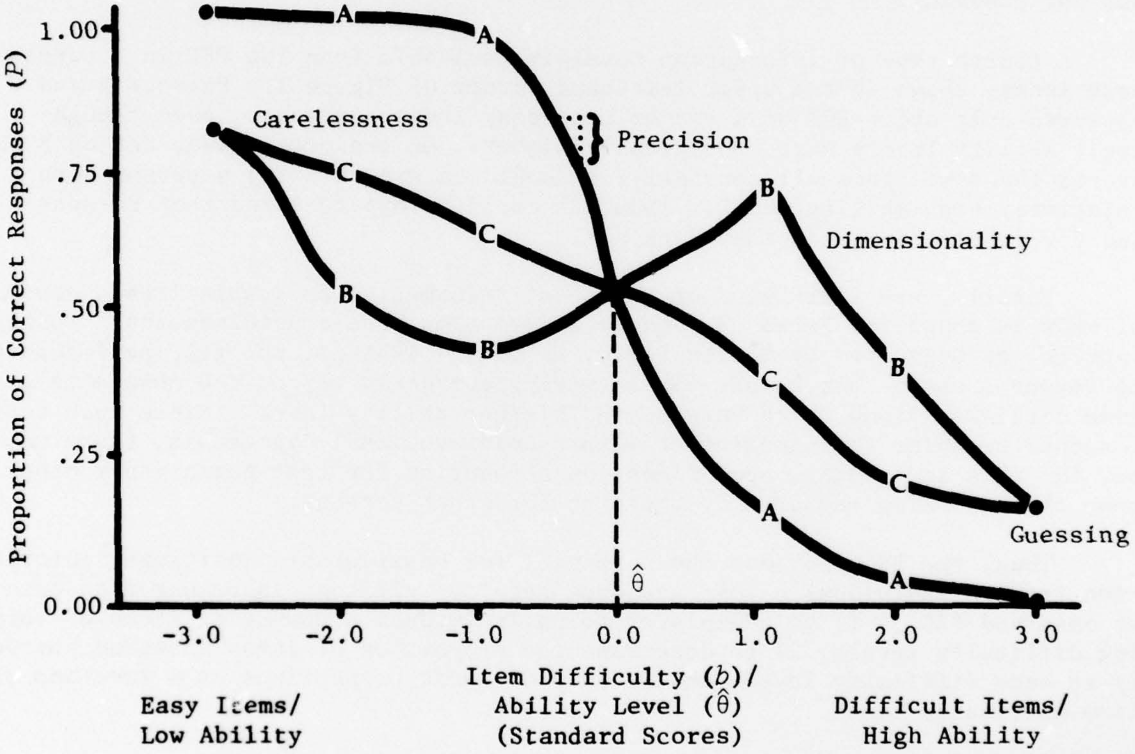


Figure 1 shows the PRC plots of three different persons--A, B, and C--who have all obtained the same score on the test by answering 50% of the total test questions correctly. Thus, all the curves cross at the point on the vertical axis of .50, and their average proportion correct across all item difficulty levels is .50. The center point of the curve can then be projected downward to the horizontal axis to obtain an ability level estimate ($\hat{\theta}$) of 0.0, which in standard score terms is at the mean of a population. Yet, Figure 1 illustrates that although these three persons all achieved the same total score on this test, they obtained that score in substantially different ways.

As shown in Figure 1, the three testees--A, B, and C--differ in a number of variables. Note that the curve for Person A has a substantially steeper slope around its center point than does that for Persons B and C. This shows that with this particular item pool, Person A was measured more precisely than either Person B or C, or (in Mosier's, 1942, terms) that the error of measurement for Person A was smaller. Thus, in addition to ability level scores, information on individual precision of measurement is derivable from the PRC.

The third type of information derivable from the study of PRCs is a person's guessing behavior. This is shown in Figure 1 as the lower right-hand portion of the curve for each testee. Note that Persons B and C correctly answered very difficult items at a nonzero level. It may, therefore, be hypothesized that they were guessing. However, Person A answered none of the difficult items correctly. It may be hypothesized that this testee, unlike the other two, was not guessing.

A fourth type of information possibly derivable from the PRC is a carelessness index, shown in the upper left-hand corner of Figure 1. Persons B and C answered only about 80% of a set of very easy items correctly, even though their ability levels were considerably higher. On the other hand, Person B answered the same items all correctly, as would be expected for a person with a relatively high ability level. Thus, it could be hypothesized that Persons B and C were more careless than Person A.

Finally, the fifth kind of potential information derivable from a study of PRCs is shown for Person B and is a deviation from a unidimensional response pattern, as suggested by Mosier (1940, p. 364). That is, the test performance of Person B shows that he/she was answering correctly beyond the chance level some difficult items which were beyond his/her ability level. Since such test response behavior is inconsistent with a unidimensional hypothesis, there may be, for this individual, some dimension accounting for test performance other than the one being measured by the test for other persons.

Thus, the PRC provides the potential for considerable additional information from an individual's test response record. All that is required to obtain an observed PRC is 1) to administer to an individual a number of items of varying difficulty levels, 2) to determine the proportion of items answered correctly at each difficulty level, and 3) to plot those proportions as a function of item difficulty level.

Expected Person Response Curves

Although the observed PRCs are useful in describing a person's test behavior, by themselves they provide no means of determining whether observed fluctuations in the curve represent important characteristics of the individual or merely chance deviations. ICC theory, however, permits the derivation of *expected* PRCs, which can then be used to evaluate whether aspects of the observed PRCs are real or chance fluctuations. In addition, these observed PRCs permit testing the fit of individual persons to the ICC model for a given set of test item responses.

Expected PRCs are derivable from either the one-, two-, or three-parameter ICC models. Derivation of the expected PRC requires an ability estimate, $\hat{\theta}$, and the item parameters for all the items administered. Generally, the ICC item parameters of the items administered will have been estimated in advance by a method such as Lord's LOGIST (Wood, Wingersky, & Lord, 1978) or one of Urry's (e.g., Schmidt & Urry, 1976) estimation procedures; the difficulty (*b*) parameters will have been used to order the items by difficulty level to obtain the observed PRC. Estimates of ability level ($\hat{\theta}$) may be obtained using programs described by Bejar and Weiss (1979).

In the case of the three-parameter logistic ICC model, the expected probability of a correct response for any given test item (P_g) is given as a function of $\hat{\theta}$, a, b, and c by the three-parameter logistic equation:

$$P_g(\hat{\theta}) = c_g + (1 - c_g) \frac{Da_g(\hat{\theta} - b_g)}{1 + e^{Da_g(\hat{\theta} - b_g)}}, \qquad [1]$$

where

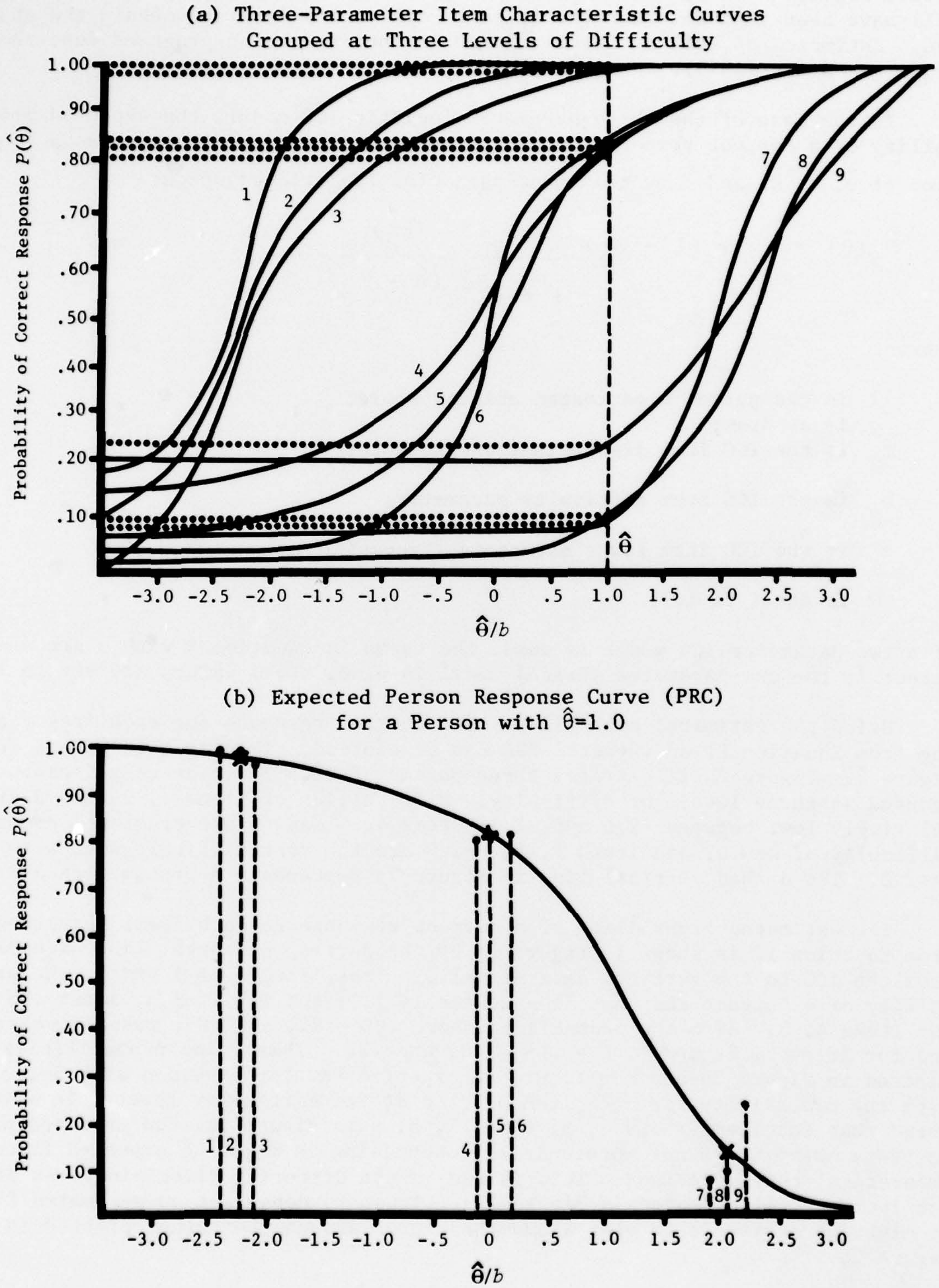
$\hat{\theta}$ is the person's estimated ability score:
g is an item;
a_g is the ICC item discrimination parameter;
b_g is the ICC item difficulty parameter;
c_g is the ICC item lower asymptote ("guessing") parameter' and
D is equal to 1.7.

If a two-parameter ICC model is used, the terms in Equation 1 with c are deleted; if the one-parameter (Rasch) model is used, the a values are set to 1.0.

Using the estimated probability of a correct response for each item resulting from Equation 1, an expected PRC can be plotted. This is illustrated in Figure 2. Figure 2a illustrates three-parameter ICCs for nine test items, grouped at three levels of difficulty. Difficulties of Items 1, 2, and 3 are relatively low, between -2.0 and -2.5; Items 4, 5, and 6 are clustered around a difficulty of b=0.0; and Items 7, 8, and 9 are the most difficult set, with $b \simeq +2.0$. The dashed vertical line in Figure 2a represents a person with a $\hat{\theta}$=1.0.

The estimated probability of a correct response to each item, resulting from Equation 1, is shown in Figure 2a by the dotted horizontal line extending from the ICC to the vertical axis at $\hat{\theta}$=1.0. Thus, for Items 1 and 2, the probability of a correct response is essentially 1.0; and for Item 3, about .98. For Items 4, 5, and 6 the probabilities are .80, .82, and .85, respectively; and for Items 7, 8, and 9, P = .08, .10, and .22. These nine probabilities are plotted in Figure 2b and constitute an expected PRC for a person with $\hat{\theta}$=1.0, with the probability for each item plotted at its difficulty level. It will be noted that for Item Groups 4, 5, 6 and 7, 8, 9 in Figure 2a, the expected proportions correct are not monotonically decreasing as might be expected from theoretical considerations. This is due to the differing discriminations of the items (as illustrated in Figure 2a). Thus, to construct an estimated PRC, it might be desirable to plot a smoothed curve around the values plotted in Figure 2b.

Figure 2
Estimating the Expected Person Response Curve (PRC) for a Person with $\hat{\theta}=1.0$ Using Nine Test Items



(a) Three-Parameter Item Characteristic Curves Grouped at Three Levels of Difficulty

(b) Expected Person Response Curve (PRC) for a Person with $\hat{\theta}=1.0$

One way of smoothing expected PRCs is to average the probabilities of a correct response to items close in difficulty level. Since the observed PRC utilizes the proportion of correct responses to a set of items of similar difficulty, averaging of the probabilities of correct responses in the expected PRC will facilitate the direct comparison of observed and expected PRCs. Lord has referred to

$$\zeta = \sum_{g=1}^{k} P_g(\theta) \qquad [2]$$

as the expected true score on a set of test items, where k is the number of items for which the expected probability of a correct response has been computed from Equation 1 and ζ is the expected number of correct responses in k items. An estimate of the proportion of correct responses on a subset of items is

$$\hat{p}_s = \zeta/k = \sum_{i=1}^{k} P_g(\theta)/k \qquad [3]$$

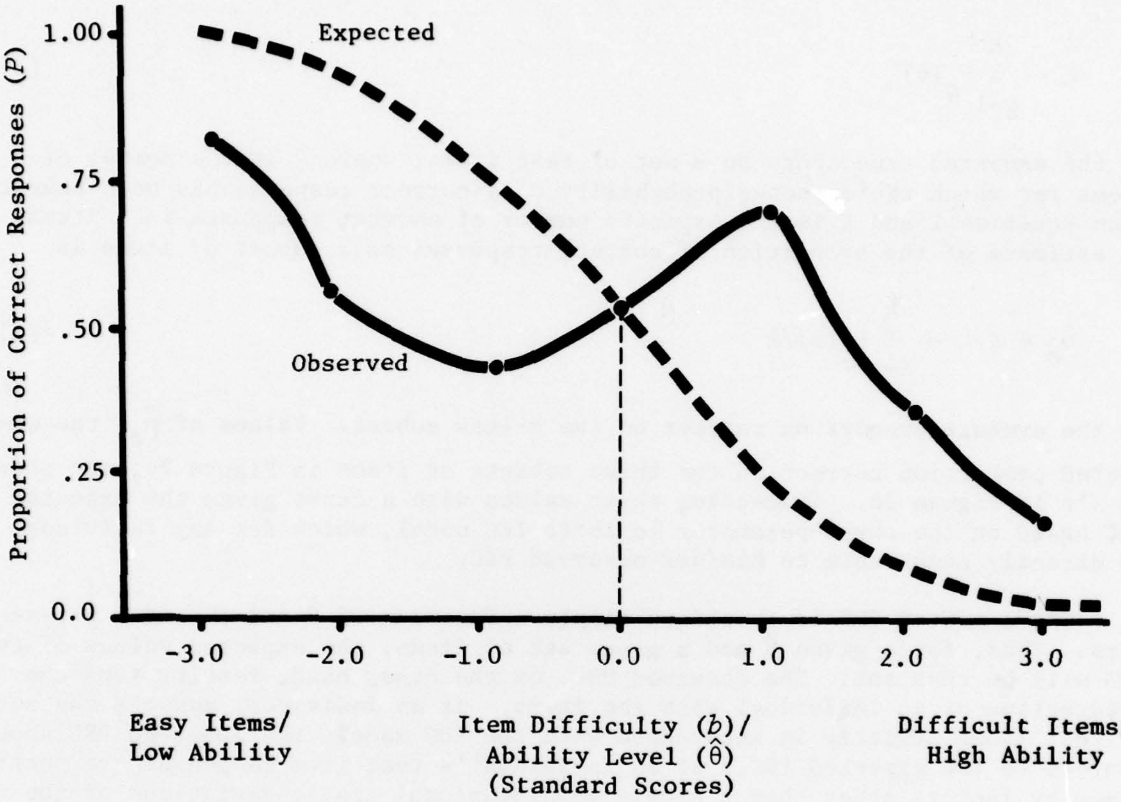
or the average proportion correct on the k-item subset. Values of $\hat{p}_s$, the expected proportion correct on the three subsets of items in Figure 2a, are shown by X's in Figure 2b. Connecting these values with a curve gives the expected PRC based on the three-parameter logistic ICC model, which for any individual is directly comparable to his/her observed PRC.

The expected PRC is therefore simply a function of $\hat{\theta}$ and the item parameters. Thus, for a given $\hat{\theta}$ and a given set of items, the expected values of the PRC will be constant. The observed PRC, on the other hand, results from the interaction of an individual with the items. If an individual answers the set of test items strictly in accordance with the ICC model, the observed PRC should conform to the expected PRC. If an individual's test item responses are determined by factors other than a single unidimensional trait, deviations of the observed PRC from the expected PRC will appear.

Observed versus Expected PRCs

Figure 3 shows hypothetical observed and expected PRCs for an individual with $\hat{\theta}$=0.0. The observed PRC (solid line) is plotted from data on test items grouped at seven points on the item difficulty continuum: b=±3, ±2, ±1, and 0. The expected PRC data points (dashed line) were derived from Equations 1 and 3 for the test items administered, using the same item difficulty groupings. To determine whether a person's carelessness, guessing, dimensionality, or precision are significantly different from those predicted by the model, an expected PRC may be determined for any person on any set of test items with estimated ICC parameters, and the observed PRC may be compared to it. If the observed PRC differs from the expected model-based prediction in any respect, the observed PRC describes a significant aspect of the person's testing behavior. Once quantified, these person-fit variables might then be usable in prediction situations to increase the accuracy of predictions made from test scores. This could be done by including additional information on guessing, carelessness, precision, and dimensionality and on other aspects of a person's test performance as reflected in the relationship of observed and expected PRCs.

Figure 3
Observed and Expected Person Response Curves for a Person with $\hat{\theta}$=0.0



Method

The following data analyses constitute a first examination of observed PRCs and their relationships with expected PRCs for a group of individuals on a test designed to permit study of the characteristics of PRCs. The major analyses were directed at establishing the reliability of observed PRCs and the fit of observed and expected PRCs. Some correlates of person-fit indices derived from the PRC were also investigated.

Subjects

Subjects were 151 undergraduate students in the introductory psychology course at the University of Minnesota. These students volunteered for the study in return for bonus points that would count toward their final grade. Students were given a posttest debriefing, which consisted of a brief explanation of the purpose of the study. No test results were given, due to the lengthy procedures for keypunching and scoring the data.

Test Instrument

The test consisted of 216 five-option multiple-choice vocabulary items. The items were chosen from a preexisting item pool of over 500 items with ICC difficulty and discrimination parameters that had been developed on a similar population of undergraduates in the introductory psychology course in previous years (McBride & Weiss, 1974). The 216 items were selected for high discriminating power and for spread of difficulty (c parameters were set at .20 for all items).

The test was given as a paper-and-pencil test without time limits. Items were randomly ordered for administration so that easy and difficult items were spread throughout the test. In addition, to control for any effects of item order, the pages of test questions were ordered in six different ways so that only one-sixth of the students took the test in the same page order.

Observed PRCs

Stratifying the test. In order to transform student response data into observed PRCs, test items were divided into strata containing an equal number of items, with each stratum representing a different level of difficulty. This was done by reordering the items by difficulty level (b parameter), then dividing them into nine separate groups (or strata) of 24 items each. In this way, Stratum 1 contained the 24 easiest items and Stratum 9 contained the 24 most difficult items.

Items were then ordered within each stratum by discrimination (a) level, with the most discriminating item the first item in the stratum and the least discriminating item the 24th item in the stratum. To investigate the parallel forms reliability of observed PRCs, each stratum was then split into two parallel substrata of items with similar difficulty and discrimination parameters. This provided 18 substrata of 12 items each. Item difficulty and discrimination parameters for all items by stratum and substratum are in Appendix Table A.

Items were scored as either correct ("1") or incorrect ("0"), with omitted items scored as incorrect. The correct-incorrect response vectors were then reordered by item difficulty level for each student. The proportion of correct responses was then computed on each of the nine strata and on each of the 18 substrata for each student, providing information for observed PRCs based on all 216 items (i.e., nine 24-item subtests of differing difficulty levels) and split-half parallel observed PRCs, each based on nine 12-item subtests.

To examine the characteristics of the items constituting the strata, internal consistency reliability of each of the nine strata was determined using Cronbach's alpha. Parallel forms reliability of the nine pairs of parallel substrata was determined by the product-moment correlation coefficient between proportion-correct scores on each of the nine pairs of substrata.

Estimated PRCs

Using Program LINDSCO (Bejar & Weiss, 1979), Owen's Bayesian ability estimation procedure was used to compute ability estimates ($\hat{\theta}$) for each student based on his/her responses to all 216 items in the test. This $\hat{\theta}$ was then used in Equations 1 and 3, in conjunction with the item parameters for the 24 items in each stratum, to obtain the expected proportion-correct score in each of the nine strata ($\hat{p}_s$). The $\hat{p}_s$ values then constituted the expected PRC for each student, assuming the three-parameter ICC model. This process was repeated for each of the parallel substrata, yielding expected PRCs for each student from each of the two 108-item parallel pools.

Correlates of Observed PRCs

In addition to the vocabulary items, 11 five-alternative Likert-type questions were used to assess psychological variables hypothesized to be related to PRC data. These questions were taken from psychological reactions scales developed by Betz and Weiss (1976), with some slight modifications. Four items were used in a Perceived Test Difficulty scale, four in a Test-Taking Anxiety scale, and three items in a Test-Taking Motivation scale.

The psychological reactions scale items (shown in Appendix Table B) were scored "1" through "5," with the first response alternative for each item scored as "1" and each succeeding alternative scored a point higher. Item scores were weighted positively or negatively (see Table B), according to how they were keyed on the psychological reactions scale. The total number of item score points ranged from +8 to -8 on the Perceived Test Difficulty and the Test-Taking Anxiety scales, and from +9 to -3 on the Test-Taking Motivation scale.

Reliability of Observed PRCs

Within- and between-persons D^2 indices. To determine the split-half parallel forms reliability of observed PRCs, a D^2 statistic was computed for each student, comparing his/her observed PRC data (proportion correct) on each of the paired substrata; thus, D^2 indexed the similarity of the two split-half PRCs for each student. A D^2 value of zero would indicate that the two split-half PRCs were identical; large values would indicate differences between the two PRCs.

Although the D^2 statistic is a commonly used descriptive statistic in comparing profiles (Cronbach & Gleser, 1953), no sampling distribution is available for it. In order to obtain some data with which to compare the split-half D^2 data, four other sets of between-persons D^2 statistics were computed for comparison purposes with the within-persons reliability D^2. Students were paired randomly into 75 pairs. The first D^2 statistic [D(AA)] was obtained by comparing the observed PRC data for one of the split-half PRCs (arbitrarily designated "A") of each individual student with those of his/her randomly paired student. The second D^2 statistic [D(BB)] was obtained by comparing the same pairs on their observed PRC data from their other (Subset B) substrata. The third and fourth D^2 statistics [D(AB) and D(BA)] were obtained by comparing one student's first split-half PRC with the other student's second split-half observed PRC.

Group means and standard deviations were then computed for the four between-persons D^2 indices [D(AA), D(BB), D(AB), and D(BA)] and the one within-persons D^2. Fisher's t test for differences in means was used to determine if within-persons split-half observed PRCs on parallel forms of the test were more similar to each other than they were to the between-persons D^2 from randomly selected individuals. If observed PRC data were reliable, it would be expected that profiles within persons would have significantly lower mean D^2 values than profiles between persons, especially considering differences in ability level between randomly paired individuals.

Chi-square tests of independence. A second approach to the study of the reliability of observed PRCs used a chi-square test of independence. For each student, the 2 × 9 contingency table included the number of correct responses on each of the parallel substrata in each of the rows of the 9-column table. Chi-square tests of independence were computed separately for each student. If the paired substrata were parallel, a nonsignificant value of chi-square would be supportive of the reliability of observed PRC data. Although this chi-square test violated the usual assumption of independence because the cell frequencies were based on the same student's responses to all the questions, it may be argued that the students' test item responses are locally independent (i.e., are independent for a given student who has a fixed value of θ) and, therefore, that the test is not inappropriate. Further study of this problem is necessary, however, in future applications of this index.

PRCs and Person-Fit

Observed versus expected PRCs. Expected PRCs were determined for each student using the method described above. To determine if students' responses to these ability test items were consistent with the three-parameter ICC model, a chi-square goodness-of-fit statistic was computed between each student's observed and expected PRC data across the nine strata. If the PRC is an adequate index of model fit, the mean chi-square for the group would be nonsignificant. On an individual level, at an .05 level of significance, chi-square goodness-of-fit values should be statistically significant for 7.55 of the 151 students by chance alone, assuming the null hypothesis of no significant deviations from person-fit. More significant chi-square values would indicate a tendency for lack of fit in these data.

When the overall level of fit in the data is substantially different from the chance expectation, it is still difficult to conclude from the overall goodness-of-fit tests that a specific individual exhibited reliable and meaningful lack of fit to the ICC model, since a certain number of such deviations from fit will occur by chance alone. To identify such individuals, two separate goodness-of-fit tests were conducted for each student using their observed and expected PRC data on each of the parallel substrata. This yielded two chi-square model fit statistics for each student--one for each of the two sets of substrata. Assuming that the two chi-square values were independent, reliable person-non-fit would be indicated by identifying persons with significant ($p < .05$) chi-square values for each of the substrata tests of independence; the probability of observing such a result by chance alone would be .05 × .05, or .0025.

PRCs and ability level. If the responses of most persons fit the ICC model, the observed PRC should be a function of ability level (θ), just as the expected PRC is a function oi ability level. To investigate this possibility, a variation of the D^2 reliability analysis was used. Based on observed PRC data within substrata, students were first matched on ability level ($\hat{\theta}$) before the between-persons D^2 measures were computed. These mean $\hat{\theta}$-matched between-persons D^2 values were then compared to the within-persons D^2 values, on the hypothesis that there should be little difference between these means (and considerably less difference than when persons were matched without regard to $\hat{\theta}$) if observed PRCs were primarily a function of ability level.

Correlates of Observed PRCs

Pearson product-moment correlations were computed among scores on the three psychological reactions scales, the within-persons D^2, and the overall person-fit chi-square. Assuming the validity of the psychological reactions scales, it would be expected that both the D^2 and chi-square values would correlate positively with Perceived Test Difficulty and Test-Taking Anxiety. Chi-square and D^2 values were also correlated with ability estimates ($\hat{\theta}$).

Results

Test Characteristics

Table 1 shows the means, standard deviations, and range of item difficulties (*b*) and proportion-correct scores (*p*) in each of the nine strata, and the values of Cronbach's alpha internal consistency coefficient for each of the 24-item strata. The strata contained items of steadily increasing difficulty: Stratum 1 contained the easiest items and Stratum 9 contained the most difficult items. This distribution of items was mirrored in the proportion-correct data for each stratum. The average proportion correct decreased as difficulty level of items increased. An exception to this tendency occurred for Strata 8 and 9, in which average proportion correct was very similar. Although average proportion correct was related to the item difficulties in accordance with expectations, the data on the range of individual proportion-correct scores shows considerable variability in proportion correct within each of the nine strata. The largest range of proportion correct was in Stratum 4 where at least one student answered only .04 of the items correctly and the maximum observed proportion correct was 1.0. The smallest range of observed proportion correct was for Stratum 9, in which the minimum proportion-correct score was .04 and the maximum was .79. These data suggest a wide range of individual differences in the proportion-correct scores for each stratum and consequently the potential for individual differences in observed PRCs.

Table 1
Mean, Standard Deviation, and Range of Item Difficulties (*b*) and Proportion-Correct Scores (*p*), and Cronbach's Alpha Coefficient for Each of the Nine Strata

	Item Difficulties (*b*)				Proportion Correct (*p*)				
			Range				Range		
Stratum	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max	Alpha
1	-2.40	.31	-2.97	-1.97	.866	.146	.130	1.000	.82
2	-1.55	.20	-1.93	-1.25	.794	.186	.210	1.000	.85
3	-1.01	.14	-1.24	-.77	.713	.216	.170	1.000	.86
4	-.56	.13	-.76	-.37	.615	.202	.040	1.000	.80
5	-.15	.11	-.36	.01	.545	.209	.080	1.000	.81
6	.26	.13	.06	.47	.481	.210	.040	.960	.80
7	.75	.19	.51	1.12	.416	.197	.080	.960	.78
8	1.32	.12	1.13	1.52	.330	.135	.040	.880	.54
9	1.98	.37	1.52	2.67	.334	.124	.040	.790	.44

Table 1 shows that the alpha internal consistency coefficients for Strata 1 through 7 were fairly high and quite similar, ranging from .78 to .86. Alpha coefficients for Strata 8 and 9 were lower--.54 and .44, respectively. The low alphas for Strata 8 and 9 were likely due to large amounts of random guessing for most students as the average porportion of correct responses of .33 for the two strata approached the theoretical expectation of .20 for the five-alternative multiple-choice items.

Table 2
Means and Standard Deviations of Item Difficulties (*b*) and Proportion-Correct Scores (*p*) in Each of the Nine Pairs of Parallel Substrata (A, B)

	Item Difficulties (*b*)				Proportion Correct (*p*)			
	Substratum				Substratum			
	A		B		A		B	
Stratum	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
1	-2.320	.296	-2.475	.317	.850	.156	.880	.166
2	-1.606	.240	-1.497	.153	.753	.202	.832	.198
3	-1.017	.161	-.993	.126	.713	.224	.711	.239
4	-.549	.137	-.572	.130	.622	.222	.606	.218
5	-.123	.084	-.180	.120	.545	.218	.543	.244
6	.296	.129	.219	.134	.460	.214	.501	.246
7	.762	.208	.740	.177	.401	.202	.429	.234
8	1.290	.113	1.354	.126	.341	.163	.317	.157
9	2.043	.411	1.910	.334	.295	.150	.370	.160

Table 2 provides data on each of the nine pairs of parallel substrata of 12 items each, including the means and standard deviations of item difficulties (*b*) and proportion-correct scores (*p*). Proportion-correct scores for each of the 151 students on each of the 18 substrata are in Appendix Table C, along with total proportion correct and the estimated ability level for each student. As Table 2 indicates, the substrata contained parallel items in the sense of similar means and standard deviations of difficulties. The smallest difference in mean difficulty was *b*=.002 for Stratum 3; the largest difference was *b*=.155 for Stratum 1, with a mean difference of .07. The proportion correct obtained by the students on the substrata were also fairly equal in mean and standard deviation. The smallest difference in mean proportion correct for the paired substrata was *p*=.002 for Stratum 3 and Stratum 5; the largest difference in mean observed proportion correct for the paired substrata was .075 (Stratum 2), indicating a high degree of similarity in mean proportion correct for the substrata.

Table 3 shows the estimated alpha coefficients for the 12-item substrata and the parallel forms correlations obtained by correlating proportion-correct scores for the 151 students on each of the nine pairs of substrata. The estimated 12-item alphas were obtained using the Spearman-Brown formula from the 24-item alphas for the strata shown in Table 1; these values were used in correcting for attenuation the parallel forms correlations. As Table 3 shows, the uncorrected parallel forms correlations between pairs of substrata ranged from .63 to .74 for the first seven strata; for the two most difficult strata the correlations were .42 and .28. These correlations were fairly substantial, considering the low internal consistency reliabilities for the two most difficult strata. Using the Spearman-Brown formula to correct the parallel forms correlations based on two 12-item tests to the 24-item length of the strata, the average corrected correlation between the pairs of Substrata 1 through 7 was slightly above .80. For the most difficult two strata, the corrected correlations were .59 and .44.

Table 3
Estimated Alpha Coefficients for 12-Item Substrata, and Parallel Forms Correlation of Proportion-Correct Scores--Uncorrected, Corrected by Spearman-Brown Formula, and Corrected for Attenuation--on Each of the Nine Pairs of Parallel Substrata

		Parallel Forms Correlation		
Stratum	Estimated 12-Item Alpha	Uncorrected	Spearman-Brown Corrected	Attenuation Corrected
1	.69	.64	.78	.93
2	.74	.74	.85	1.00
3	.75	.73	.84	.97
4	.67	.70	.82	1.04
5	.68	.64	.78	.94
6	.67	.66	.80	.99
7	.64	.63	.77	.98
8	.40	.42	.59	1.00
9	.28	.28	.44	1.00

To determine whether scores on the paired substrata correlated as highly as possible, given the reliabilities of the substrata, the estimated 12-item alphas for the substrata were used along with the uncorrected parallel forms correlation to estimate the correlation between proportion-correct scores on the paired substrata, assuming that the substrata had been perfectly reliable. These attenuation-corrected correlations are shown as the last column in Table 3. As the data show, attenuation-corrected correlations were .97 or above for seven of the nine strata; for Strata 1 and 5, these correlations were .93 and .94, respectively. These data indicate that the paired substrata scores were as parallel as possible, given their estimated internal consistencies.

Reliability of Observed PRCs

Within- and between-persons D^2 indices. Table 4 shows summary statistics for the within-persons D^2 on the parallel substrata and the between-persons D^2 using randomly paired individuals. The within-persons D^2 mean of .28, with a standard deviation of .15 and range of .02 to .86, were all relatively small. These data indicate that for the within-persons D^2, the average difference in proportion correct on the paired substrata was about p=.18. By comparison, the between-persons D^2 mean was .75, with a standard deviation of .66 and a range of .07 to 4.09. Thus, the average difference in proportion correct between randomly paired individuals was about p=.29.

Table 4
Mean, Standard Deviation, and Range of Within- and Between-Persons D^2 Indices, and Results of *t* Tests Comparing the Mean Within-Persons D^2 with Each Between-Persons D^2 Index

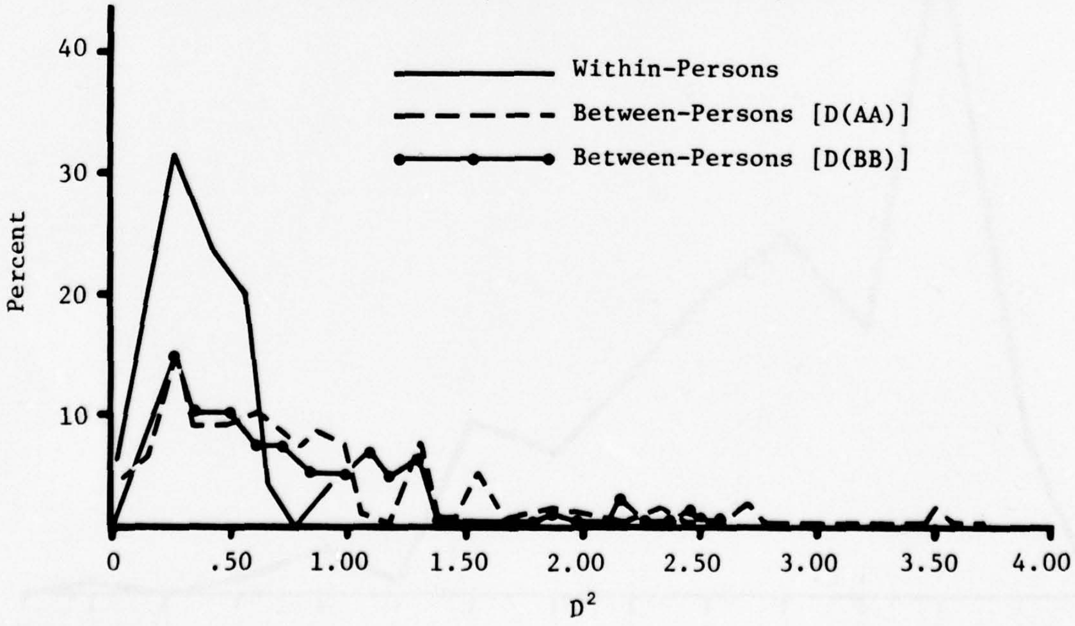
				Range			
D^2 Index	*N*	Mean	*SD*	Min	Max	*t*	*p**
Within-Persons	150	.28	.15	.02	.86		
Between-Persons							
D(AA)	75	.70	.64	.08	3.92	7.64	<.001
D(BB)	75	.78	.68	.06	4.32	8.62	<.001
D(AB)	75	.76	.69	.11	4.50	8.14	<.001
D(BA)	75	.76	.62	.04	3.60	9.06	<.001

*Probability of error in rejecting null hypothesis of no difference in group means.

The smaller within-persons D^2 demonstrates greater split-half profile similarity within persons than between pairs of randomly selected persons, irrespective of which split-half test was used for the between-persons comparisons. The *t*-test statistics in Table 4 demonstrate this sizable difference between the two types of profile comparison. Although the *t*-test assumption of independent groups was violated in these data, the mean differences in each case were substantial enough to support the conclusion that the PRCs are reliable.

Figure 4 provides further data on the distribution of the D^2 indices in terms of the relative frequency distributions of the within-persons reliability D^2 and two of the between-persons D^2 indices [D(AA) and D(BB)]. As Figure

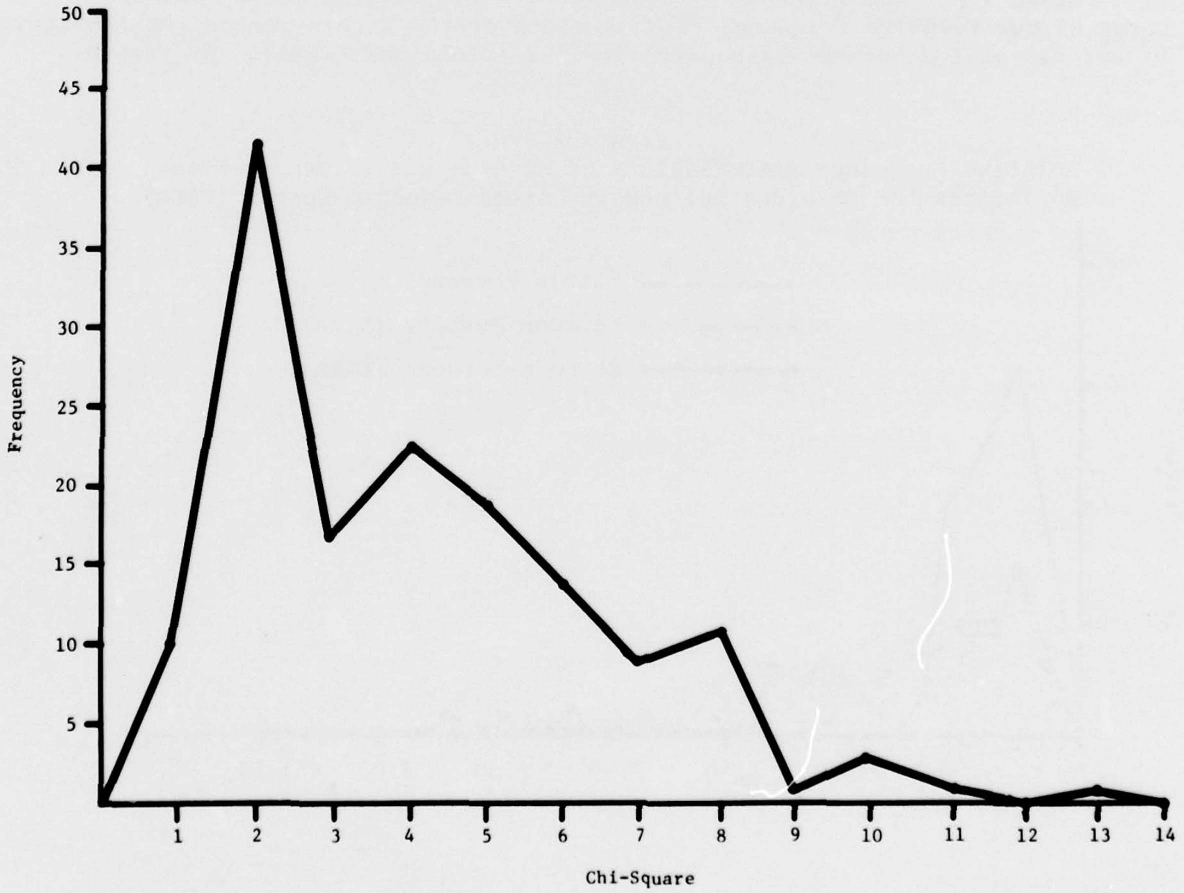
Figure 4
Relative Frequency Distributions of Within- and Between-Persons D^2 Indices for Observed Split-Half Person Response Curves (PRCs)



4 shows, there was little overlap between the two distributions. Virtually all of the within-persons D^2 values were below .75; and the distribution was highly peaked with a modal value very close to zero, indicating that the observed PRCs from the split substrata were very similar for most of the 150 students. By contrast, although the mode of between-persons D^2 indices was similar to that of the within-persons D^2, the relative frequency associated with that mode was considerably less than that of the within-persons distribution, and the distributions of the between-persons data was considerably less peaked.

Chi-square tests of independence. Results of the chi-square test of independence, based on a 2 × 9 contingency table with number-correct scores on each of the nine pairs of parallel substrata for each student, are shown in Figure 5. The minimum value of chi-square was .14 and the maximum was 12.94; mean of the distribution was 3.67, with a standard deviation of 2.34. A chi-square value of 15.51 is statistically significant at the p=.05 level with 8 degrees of freedom (from the 2 × 9 contingency table). Since all of the individual chi-square values were less than 15.51, the data show that the two split-pool observed PRCs for all students were not significantly different from each other, further supporting the D^2 data which indicated that the observed PRCs obtained from these data were reliable.

Figure 5
Frequency Polygon of Rounded Intra-Individual
Chi-Square Test of Independence Values Between Person
Response Curves (PRCs) for the Nine Pairs of Parallel Substrata

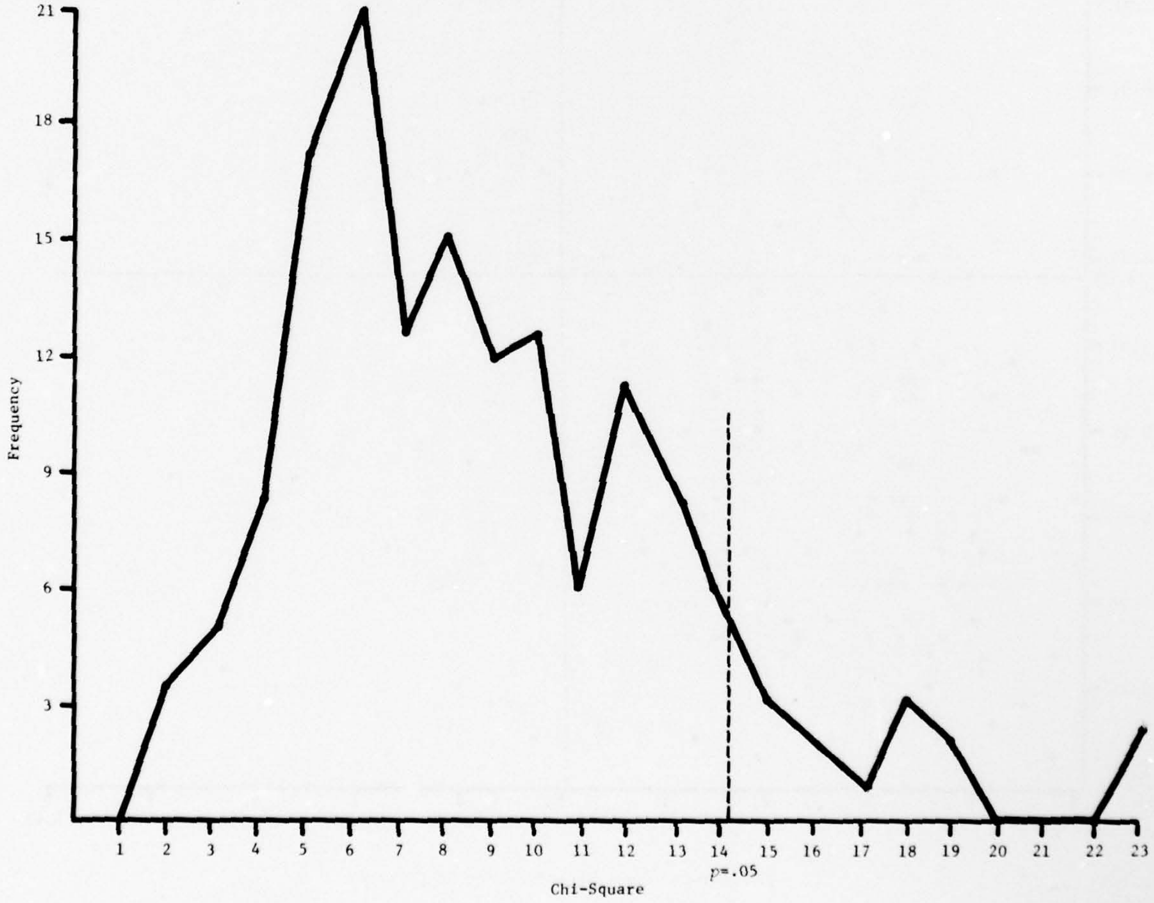


PRCs and Person-Fit

The frequency distribution of individual chi-square values reflecting the fit of observed PRCs to the PRCs expected from the three-parameter ICC model is shown in Figure 6. The lowest chi-square value obtained was 1.88 and the highest was 23.17. Mean chi-square was 8.76, with a standard deviation of 4.14; modal value was about 6.0.

Since ability estimates used in calculating the theoretically expected proportion-correct scores were taken from the data being analyzed, an extra degree of freedom was subtracted to determine the significance of the chi-square values. Thus, with 7 degrees of freedom, a chi-square value of 14.07 is significant at the .05 level. The group mean chi-square was well below this value, which would suggest that the three-parameter logistic ICC model served as a fairly good predictor of test response behavior for the majority of this group of students. Of 151 students, 8 would be expected to have significant chi-square values by chance alone at the .05 level; in this group, 15 students had chi-square values greater than 14.07.

Figure 6
Frequency Distribution of Intra-Individual Chi-Square
Values for Goodness of Fit Between Observed and
Expected Person Response Curves (PRCs)



To identify persons reliably deviating from the model, the chi-square person-fit statistics were recomputed for each student separately on the two sets of substrata. The joint distribution of chi-square values for the 151 students is shown in Figure 7, with the .05 significance level indicated by the dashed horizontal and vertical lines. Persons in the upper right-hand quadrant were identified as those deviating significantly from the expected values, with *p*=.0025. As Figure 7 shows, six students had significant chi-square values for both pairs of substrata and were thus placed in the upper right-hand quadrant. Of these six, four were also significantly non-fitting on the overall chi-square goodness-of fit test. These four are indicated in Figure 7 by their subject numbers, and their PRCs (both observed and expected) are in Figure 9. Persons 83, 111, 138, and 117 might be hypothesized to have reliably non-fitting PRCs. Of the 15 students whose overall chi-square values were statistically significant, those not included in the upper right-hand quadrant may be hypothesized to be non-fitting only by chance.

Figure 7
Joint Distribution of Intra-Individual Chi-Square Values for Goodness of Fit for Odd- and Even-Numbered Substrata

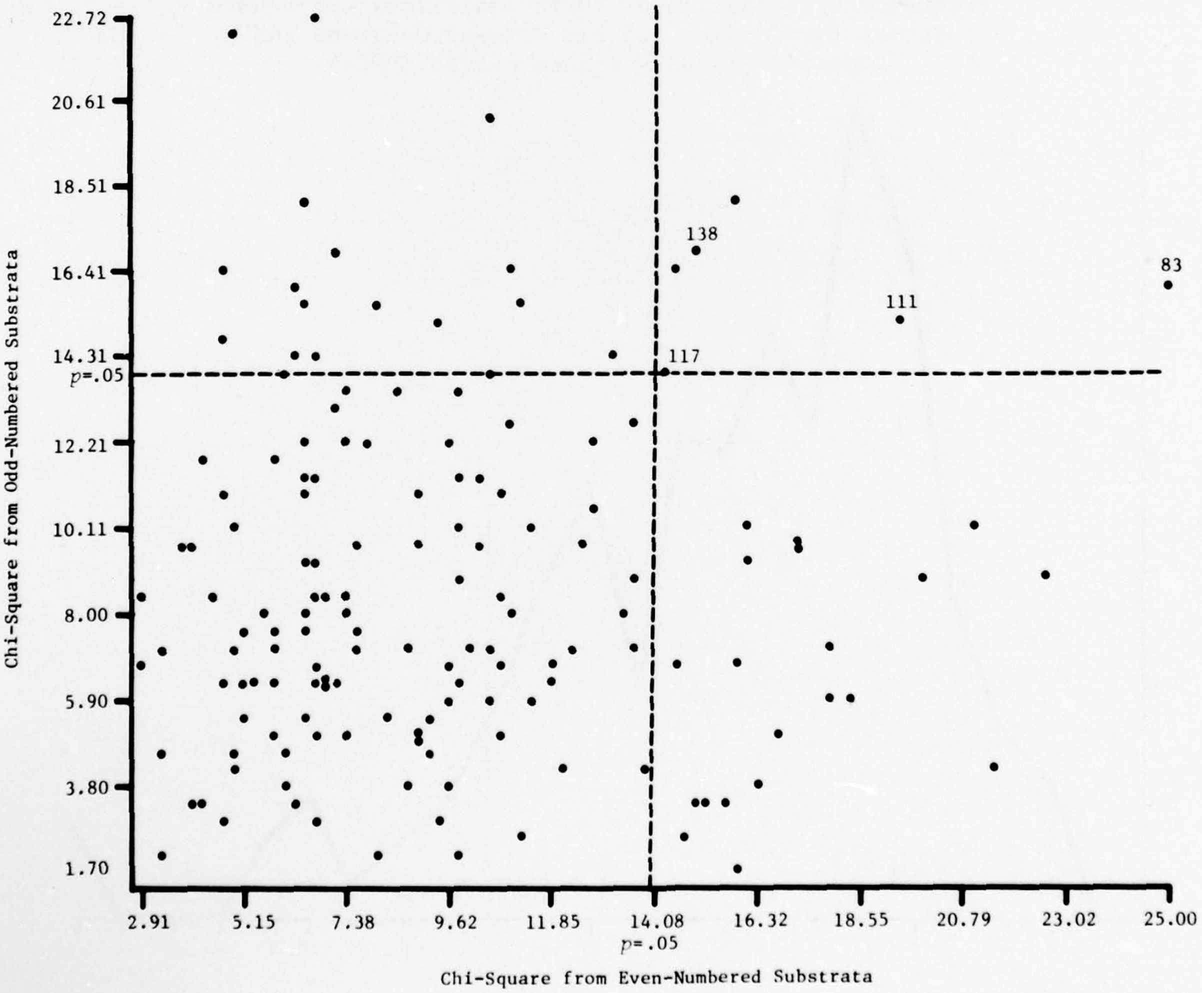


Figure 8 shows observed and expected PRCs for students with low overall chi-square person-fit values. Person 128 (Figure 8a) obtained the lowest chi-square value among the 151 students tested. As Figure 8a shows, the observed PRC for Person 128 (solid line) was quite close to the expected PRC (dashed line) for each of the nine strata. Figures 8b through 8d show expected and observed PRCs for three other students for whom model-fit was quite good, as indicated by the low chi-square values, although as expected, some minor deviations from model-fit appeared (e.g., Figure 8d) as chi-square values increased.

Figure 9 shows PRC person-fit results for four of the persons identified in Figure 7 as not reliably fitting the ICC model; these data are based on their total PRCs. The ways in which these four students' response curves deviated from their expected curves differed widely. Person 111 (Figure 9a) seems to have been careless with easier items, as indicated by a proportion correct of .75 on items in Stratum 1, and then to have been fortunate in guessing on some of the more difficult ones (*p*=.50 on Stratum 7). On the other hand this may be the type of profile to be expected from a person with an unusual educational history, such as an international student with a specialized knowledge of English. Person 117 (Figure 9b) and Person 138 (Figure 9d) seemed to have done much better on difficult items than was predicted by the model; these students might be sophisticated at guessing or high in "testwiseness." Person 83 (Figure 9c) seems to have exhibited carelessness on the easier items (Stratum 1) but more effort (with perhaps some good guesses) on the more difficult items in Strata 6 through 8.

Although these figures demonstrate lack of fit of individuals to the model-based predictions, they do not by themselves point to clear interpretations. However, they do illustrate some of the different ways in which significant deviations in test data can occur. This demonstrates the need for methods of assessing and interpreting the many ways in which non-fitting PRCs may occur.

PRCs and Ability Level

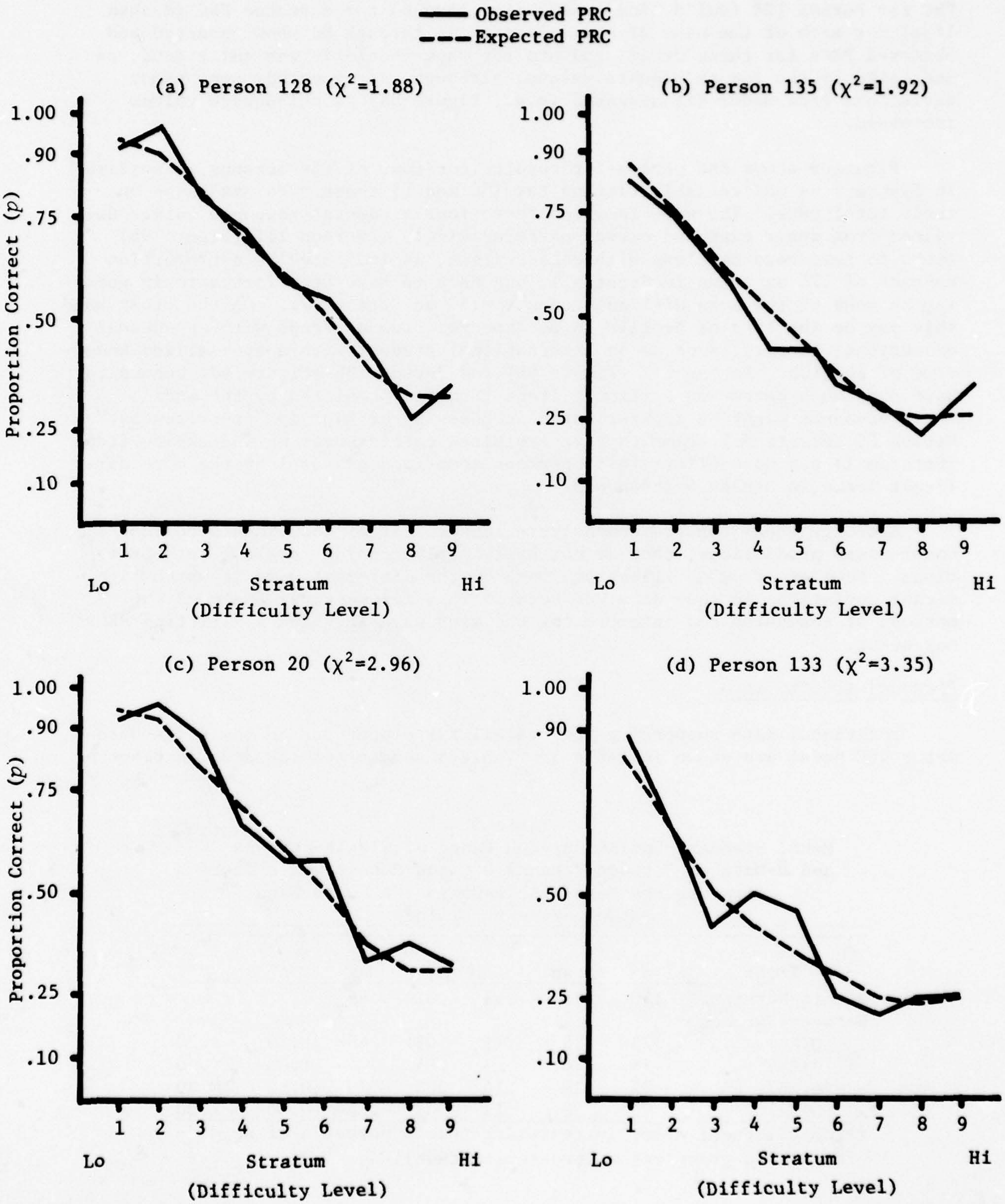
Additional data supporting the overall fit of persons to the three-parameter ICC model are shown in Table 5. Table 5 summarizes the distributions of

Table 5
Mean, Standard Deviation, and Range of Within-Persons D^2 and $\hat{\theta}$-Matched Between-Persons D^2, and Results of t Tests Comparing the Mean Within-Persons D^2 with Each Between-Persons D^2 Index

				Range			
D^2 Index	*N*	Mean	*SD*	Min	Max	*t*	*p**
Within-Persons	150	.28	.15	.02	.86		
Between-Persons							
D(AA)	75	.25	.11	.03	.48	1.50	<.20
D(BB)	75	.26	.13	.05	.74	1.00	<.50
D(AB)	75	.29	.14	.05	.79	0.50	<.80
D(BA)	75	.28	.12	.07	.64	0.00	1.00

*Probability of error in rejecting null hypothesis of no difference in group means (two-tailed test).

Figure 8
Observed and Expected Person Response Curves (PRCs) for Four Persons Whose Responses Reliably Fit the Three-Parameter ICC Model



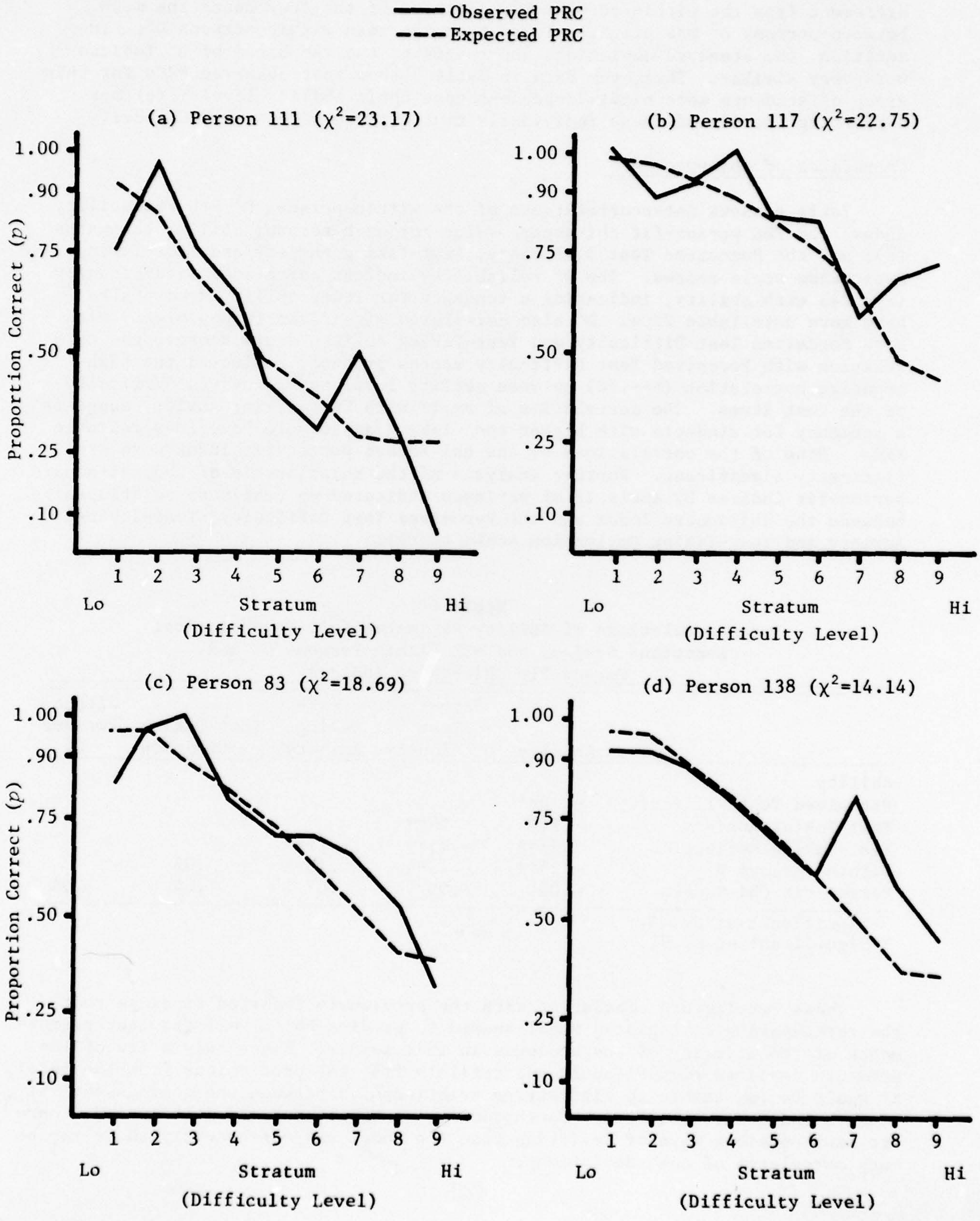
(a) Person 128 (χ^2=1.88)

(b) Person 135 (χ^2=1.92)

(c) Person 20 (χ^2=2.96)

(d) Person 133 (χ^2=3.35)

Figure 9
Observed and Expected Person Response Curves (PRCs) for Four Persons Whose Responses Did Not Reliably Fit the Three-Parameter ICC Model



(a) Person 111 (χ^2=23.17)

(b) Person 117 (χ^2=22.75)

(c) Person 83 (χ^2=18.69)

(d) Person 138 (χ^2=14.14)

between-persons D^2 data on parallel substrata when students were matched as closely as possible for $\hat{\theta}$ values before the substrata D^2 indices were calculated. As Table 5 shows, none of the mean between-persons D^2 indices was significantly different from the within-persons D^2; in three of the four cases the mean between-persons D^2 was slightly lower than the mean within-persons D^2. In addition, the standard deviations and ranges of the two kinds of D^2 indices were very similar. Thus, the data in Table 5 show that observed PRCs for this group of students were highly dependent upon their ability levels, further supporting the fit of these individuals to the three-parameter ICC model.

Correlates of Observed PRCs

Table 6 shows intercorrelations of the within-persons D^2 PRC reliability index; the PRC person-fit chi-square value for each person; ability estimates ($\hat{\theta}$); and the Perceived Test Difficulty, Test-Taking Anxiety, and Test-Taking Motivation scale scores. The D^2 reliability indices correlated significantly (r=-.24) with ability, indicating a tendency for lower ability students to have more unreliable PRCs. D^2 also correlated significantly positively with both Perceived Test Difficulty and Test-Taking Anxiety scale scores; the correlation with Perceived Test Difficulty scores probably reflected the high negative correlation (r=-.70) between ability level and perceived difficulty of the test items. The correlation of r=.18 with Test-Taking Anxiety suggests a tendency for students with higher test-taking anxiety to have less reliable PRCs. None of the correlations of the chi-square person-fit index were statistically significant. Further analysis of the relationship of the chi-square person-fit indices by analysis of variance indicated no nonlinear relationships between the chi-square index and the Perceived Test Difficulty, Test-Taking Anxiety, and Test-Taking Motivation scale scores.

Table 6
Intercorrelations of Ability Estimates ($\hat{\theta}$), Psychological Reactions Scales, and PRC Within-Persons D^2 and Person-Fit Chi-Square Indices

	Ability	Perceived Test Difficulty	Test-Taking Anxiety	Test-Taking Motivation	Within-Persons D^2
Ability					
Perceived Test Difficulty	-.70**				
Test-Taking Anxiety	-.16*	.28**			
Test-Taking Motivation	.37**	-.35**	.11**		
Within-Persons D^2	-.24**	.18**	.18**	.03	
Person-Fit Chi-Square	-.06	-.05	.07	.04	-.04

*Significant at p<.05.
**Significant at p<.01.

These results are consistent with the previously reported findings that the three-parameter logistic model seemed to predict quite well the test performance of the majority of the students in this sample. Since only a few of the students deviated significantly and reliably from the predictions from the model, it would be impossible to find strong relationships between the goodness-of-fit results and other variables. Furthermore, as was illustrated in Figure 9, there are many possible ways of deviating from the model and, consequently, there may be many correlates of such deviations.

Conclusions and Directions for Future Research

The feasibility of the person response curve (PRC) approach to investigating the fit of persons to the three-parameter ICC model was explored in this study. To operationalize the PRC it was necessary to subdivide ability test items into separate strata of varying difficulty levels. For the vocabulary test used in this study, strata possessed sufficient internal consistency and parallel forms reliability to justify their use, although the more difficult strata were much less reliable than the easier strata.

Conclusions

The PRCs proved to be highly reliable. The D^2 analyses indicated not only that intra-individual profiles were more similar than profiles between randomly selected persons but also that profiles between people of similar ability level were also very similar. As additional evidence of profile reliability, chi-square tests for independence between profiles of parallel forms for each individual were nonsignificant for all 151 students. The high correlation of r=.82 (p<.001) between the intra-individual parallel forms chi-squares and the D^2 suggests that the chi-square test may be sufficient in future studies, since it also provides a more ready means of assessing statistical significance.

The results of the D^2 statistics between individuals matched on ability level were interesting, since they illustrated close profile similarity between different persons of similar ability level. This suggests that for the majority of this sample, PRCs were predictable as a function of ability level. A more complete test of this hypothesis was conducted with a chi-square goodness-of-fit test between observed proportion-correct scores on each of nine strata and expected proportion-correct scores predicted by the three-parameter logistic model. The nonsignificant group mean suggests that the model was a reasonable way of describing students' test response behavior.

At the .05 level, eight students were expected to have significant chi-square goodness-of-fit values for observed and expected PRCs. Fifteen students had significant chi-square values, leaving somewhat in question whether these students deviated from the model because of chance or interaction with another dimension. One method of investigating this question was to calculate separately the goodness of fit of each student's observed and expected PRCs on the odd-numbered substrata and on the even-numbered substrata. Of the 15 students with significant chi-square values on the overall nine-strata goodness-of-fit test, four had significant chi-square values on both substrata goodness-of-fit tests. These four students were identified as reliably deviating from the ICC model predictions. The nature of this lack of fit, however, would best be investigated in a future study with an experimental design that included interactions with additional dimensions other than the ability being measured.

Having demonstrated the goodness of fit of observed PRCs with model-predicted PRCs, and with no firm evidence to suggest that significant results for a majority of the students were due to anything other than chance, the nonsignificant results for the relationship of the goodness-of-fit chi-square variable with nontest variables seems to follow. Scores on the psychological reactions scales correlated with each other and with ability estimates in expected ways but did not correlate significantly with the overall chi-square variable. These results substantiated the fit of the model to observed student test-response behavior. The psychological reactions scales could be used in a future study of non-fit in which these psychological states could be experimentally induced.

The results of this study demonstrate that the PRC can be useful in studying the fit of individuals to ICC models by testing the fit of the observed PRC to the theoretically expected PRC. Although the three-parameter ICC model was used here, the method can be used with the two-parameter or one-parameter logistic (Rasch) model or with any of the normal ogive ICC models. The data also demonstrated that the three-parameter ICC model adequately accounted for the test response behavior of the vast majority of the students studied. More research is, of course, necessary to further explore the use of the PRC in examining model-fit in test behavior.

Directions for Future Research

Guessing and "testwiseness" are variables which are unrelated to abilities but may affect ability test scores. To determine whether these variables can be detected by PRCs or PRC-fit to theoretical predictions, a useful experiment would be to administer a multiple-choice ability test along with testwiseness and guessing scales to groups of students. One subgroup in the experimental design should be an experimental group trained in testwiseness and/or in guessing skills. The effects of testwiseness or guessing would be studied by analysis of the chi-square goodness-of-fit statistics comparing the expected and observed PRCs for the experimental and control groups. Special attention should be given to chi-square values on the most difficult items in the ability test rather than overall chi-squares, since it is on these items that the experimental effect is likely to be observed.

Cultural bias is another dimension which may differentially affect ability test performance (e.g., Church, Pine, & Weiss, 1978; Martin, Pine, & Weiss, 1978; Pine & Weiss, 1978). One approach to testing for the existence of such bias by use of PRCs would be to compare the goodness of fit of observed and expected PRCs for a control group of white middle-class testees and a group of testees who would be hypothesized to have uneven educational development by white middle-class American standards. This latter group might involve international students with a specialized knowledge of the English language or some American minority group persons. It would be expected that the PRCs would show greater deviation from the model predictions for the latter group, particularly in terms of deviations from the unidimensionality required by the ICC model.

Carelessness and nervousness are two other dimensions which may contribute to unexpected performance on ability tests and which may be detected by PRC analysis. To study the effect of these dimensions on person-fit, an ability test could be administered to three groups of randomly selected individuals from the same population. A low-motivation-possibly-careless control group would be given minimal information about the test. Treatment Group 1 would be told that the test results did not matter and that the experimenter just needed to fill his/her quota of subjects. Treatment Group 2 would be told that the test is an important determiner of whether or not they would be able to complete college or to succeed in some occupation; this would be considered the high-anxiety group. The experimentally induced states should be verified with improved versions of the psychological reactions scales for motivation and anxiety used in this report. Values comparing chi-square observed versus expected PRCs would be compared, with special attention to the PRC-fit data on the easier strata for the low-motivation group and on the more difficult strata for the high-anxiety group. This would give information on possible psychological correlates of fit on a stratum-by-stratum basis. Data of this type might be used, for example, to investigate the operation of the Yerkes-Dodson Law (Taylor & Spence, 1958; Yerkes & Dodson, 1908) in ability test data; PRC-fit data would support this hypothesis if high-anxiety testees perform better than expected on easy test items and more poorly than expected on the more difficult test items.

Further investigation of the measurement properties of observed versus expected chi-square goodness-of-fit statistics for assessing non-fit of persons is also of importance. Monte carlo simulations should be run in order to determine the null distribution of the chi-square values. These should be repeated at a number of theta levels to determine whether goodness-of-fit distributions differed as a function of ability level.

Finally, since the research literature on methods for assessing non-fitting profiles has begun to branch in several different directions, it would be informative and useful to compare the efficacy of several different methods using the same data base. The one-, two-, and three-parameter ICC models could each be used in computing ability estimates so that non-fit measures based on these different models could be used. This would best be done in simulation, with non-fitting data experimentally induced so that the different methods of evaluating model-fit could be compared on their degree of "hits" and "misses."

These are only a few of many research possibilities in investigating the properties and the diagnostic utility of PRCs. A closer look at these properties of the PRC test performance profiles and their use in determining person-fit may provide important information on selected individuals and improve the validity of ability tests for individual prediction and diagnosis.

References

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1979. (NTIS No. AD A067752)

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977. (NTIS No. AD A044828)

Betz, N., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027170)

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A046062)

Church, A. T., Pine, S. M., & Weiss, D. J. A comparison of levels and dimensions of performance in black and white groups on tests of vocabulary, mathematics, and spatial ability (Research Report 78-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1978. (NTIS No. AD A062797)

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Hambleton, R., & Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Kingsbury, G. G., & Weiss, D. J. Relationships among achievement level estimates from three item characteristic curve scoring methods (Research Report 79-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1979. (NTIS No. AD A069815) (a)

Kingsbury, G. G., & Weiss, D. J. An adaptive testing strategy for mastery decisions (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1979. (b)

Levine, M. Psychometric models and appropriateness measurement. In D. J. Weiss (Ed.), Proceedings of the 1979 conference on computerized adaptive testing, in preparation.

Levine, M., & Rubin, D. Measuring the appropriateness of multiple-choice test scores (Research Bulletin 76-31). Princeton, NJ: Educational Testing Service, December 1976.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M., & Novick, M. Statistical theories of mental test scores. Reading, MA: Addison & Wesley, 1968.

Lumsden, J. Person reliability. Applied Psychological Measurement, 1977, 1, 477-482.

Lumsden, J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978, 31, 19-26.

Martin, J. T., Pine, S. M., & Weiss, D. J. An item bias investigation of a standardized aptitude test (Research Report 78-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1978. (NTIS No. AD A064352)

McBride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement (Research Report 74-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1974. (NTIS No. AD 781894)

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Mead, R. J. Using the Rasch model to identify person-based measurement disturbances. In D. J. Weiss (Ed.), Proceedings of the 1979 conference on computerized adaptive testing, in preparation.

Mosier, C. I. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. Psychological Review, 1940, 47, 355-366.

Mosier, C. I. Psychophysics and mental test theory II: The constant process. Psychological Review, 1942, 48, 235-249.

Pine, S. M., & Weiss, D. J. A comparison of the fairness of adaptive and conventional testing strategies (Research Report 78-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, August 1978. (NTIS No. AD A059436)

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. In D. J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Schmidt, F., & Urry, V. W. Item parameterization procedures for the future. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Taylor, J. A., & Spence, K. W. The relationship of anxiety level to performance in serial learning. Journal of Experimental Psychology, 1958, 57, 55-60.

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (NTIS No. AD A018758)

Vale, C. D., & Weiss, D. J. A comparison of information functions of multiple-choice and free-response vocabulary items (Research Report 77-2). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1977.

Wainer, H., & Wright, B. D. Robust estimation of ability in the Rasch model. In D. J. Weiss (Ed.), Proceedings of the 1979 conference on computerized adaptive testing, in preparation.

Weiss, D. J. Canonical correlation analysis in counseling psychology research. Journal of Counseling Psychology, 1972, 19, 241-252.

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

Weiss, D. J. (Ed.). Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975. (NTIS No. AD A018675)

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service, 1976. (modified January 1978)

Wright, B. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-115.

Wright, B., & Mead, R. J. The use of measurement models in the definition and application of social science variables (Technical Report DAHC19-76-G-011). Arlington, VA: U.S. Army Research Institute for the Behavioral Sciences, 1977.

Wright, B. D., & Stone, M. H. Best test design. Chicago: MESA Press, 1979.

Yerkes, R. M., & Dodson, J. D. The relation of strength to stimulus to rapidity of habit formation. Journal of Comparative Neurological Psychology, 1908, 18, 459-482.

Table A

Item Numbers, Discrimination (*a*) and Difficulty (*b*) Parameters, and Substratum Designation for Items in the Vocabulary Item Pool (*c*=.20 for All Items)

Item Number	*a*	*b*	Sub-stratum
Stratum 9			
343	.264	2.671	A
580	.315	2.639	B
243	.316	2.607	A
228	2.935	2.411	A
569	.400	2.402	A
394	.297	2.398	B
167	.416	2.155	A
533	.632	2.153	B
247	.647	2.063	A
577	.613	2.004	B
374	.557	1.992	A
531	.346	1.921	B
260	.709	1.820	B
504	.635	1.808	A
603	.380	1.800	B
616	.610	1.759	A
119	.534	1.729	B
521	.753	1.696	B
400	.929	1.682	B
242	.524	1.574	B
610	.788	1.566	A
159	.768	1.559	A
168	.913	1.553	B
350	.317	1.525	A
Stratum 8			
383	2.111	1.518	B
525	.570	1.509	B
147	.825	1.469	A
253	2.321	1.443	B
572	1.289	1.433	B
213	.429	1.430	A
368	.462	1.424	B
216	.668	1.397	B
217	1.249	1.384	A
660	.829	1.369	B
291	1.641	1.354	A
333	.351	1.340	B
397	.651	1.339	A
586	1.536	1.309	A
268	.270	1.300	A
259	.365	1.293	B
341	.634	1.282	A
581	1.256	1.207	A
306	1.317	1.204	B
231	.874	1.186	B
617	2.778	1.172	A
164	.687	1.136	B
238	.758	1.130	A
576	.427	1.128	A
Stratum 7			
516	.350	1.116	B
601	1.315	1.097	A
215	.908	1.069	A
111	.822	.936	A
375	.832	.934	B
526	1.169	.919	A
Stratum 7, *cont'd.*.			
523	1.210	.875	B
302	.845	.846	B
271	.886	.796	A
139	.614	.794	B
324	.524	.772	A
267	.650	.770	A
289	.480	.691	B
113	1.057	.678	B
340	1.921	.645	B
60	1.232	.643	B
590	.538	.617	A
59	1.093	.601	B
372	.346	.559	A
593	.560	.551	B
264	2.276	.549	A
265	1.571	.546	A
538	1.181	.518	A
266	2.120	.509	B
Stratum 6			
252	.420	.472	B
633	.712	.470	A
301	1.376	.468	A
519	.527	.440	A
377	.585	.393	B
582	1.200	.351	A
549	.433	.348	A
551	.896	.336	B
116	.494	.334	B
50	.694	.321	A
318	.526	.310	A
272	1.960	.223	A
502	.730	.218	B
52	.844	.205	A
54	.378	.204	B
622	.444	.201	A
599	1.634	.158	B
354	.327	.151	A
56	1.109	.135	B
161	1.384	.132	B
355	.506	.104	B
145	.791	.086	B
209	.870	.067	A
444	.621	.059	B
Stratum 5			
292	.610	.012	B
597	.624	-.000	A
382	.856	-.010	A
205	.603	-.024	B
207	.793	-.035	A
137	.499	-.056	A
503	1.062	-.090	B
365	.877	-.105	A
176	.415	-.106	B
154	.872	-.124	B
218	.407	-1.25	B
234	.650	-.132	A
Stratum 5, *cont'd.*			
270	1.223	-.138	A
143	1.036	-.153	A
156	.841	-.166	A
643	.487	-.202	A
211	.773	-.236	A
37	.860	-.236	B
157	.384	-.245	A
390	.797	-.257	B
224	.679	-.257	B
221	.822	-.278	B
307	.699	-.325	B
128	1.074	-.355	B
Stratum 4			
535	.767	-.374	A
58	.587	-.380	A
203	.820	-.384	B
33	.800	-.390	B
332	.973	-.396	A
130	.949	-.439	A
183	.728	-.452	B
588	.465	-.464	B
53	.637	-.478	B
222	.652	-.499	A
142	.314	-.536	A
123	.823	-.559	A
136	.317	-.562	B
293	.669	-.567	A
287	.523	-.652	A
117	.619	-.656	B
85	.934	-.670	B
584	.758	-.677	A
185	.682	-.684	B
109	1.109	-.701	B
515	1.084	-.708	B
239	.939	-.709	B
204	.876	-.742	A
87	1.241	-.763	A
Stratum 3			
112	.614	-.775	A
235	.664	-.776	A
36	1.644	-.789	B
546	.555	-.801	B
615	.439	-.858	B
43	1.108	-.861	A
371	.444	-.916	B
194	1.790	-.959	A
47	1.043	-.962	A
103	1.059	-.999	B
26	.364	-1.020	A
285	.835	-1.022	A
637	.877	-1.023	B
40	1.236	-1.032	A
51	1.432	-1.043	B
241	.568	-1.054	B
173	.882	-1.062	B
322	.673	-1.091	B
Stratum 3, *cont'd.*			
199	1.093	-1.093	B
108	.536	-1.155	A
86	.887	-1.189	A
189	.757	-1.191	B
141	.478	-1.208	A
227	.812	-1.245	A
Stratum 2			
232	.673	-1.251	B
191	1.749	-1.257	A
88	.706	-1.332	B
186	1.067	-1.335	A
127	1.075	-1.345	A
129	1.274	-1.352	B
101	1.165	-1.395	A
44	1.145	-1.412	B
311	.746	-1.430	A
190	1.818	-1.439	B
83	.875	-1.449	B
214	.476	-1.488	B
13	1.888	-1.553	B
34	.830	-1.582	B
84	1.701	-1.640	A
559	.616	-1.675	A
27	1.427	-1.675	B
95	.563	-1.707	B
96	1.129	-1.722	B
76	.618	-1.750	A
196	2.128	-1.791	A
197	.253	-1.850	A
125	1.236	-1.875	A
262	.768	-1.928	A
Stratum 1			
22	1.200	-1.971	B
158	1.083	-1.996	A
106	.672	-2.009	A
138	1.728	-2.023	A
31	.722	-2.141	B
63	.692	-2.144	A
202	.620	-2.172	B
206	1.105	-2.187	A
184	.726	-2.193	A
9	1.452	-2.240	B
80	.859	-2.251	B
126	.956	-2.266	A
602	.255	-2.285	A
68	1.014	-2.479	A
198	.801	-2.503	B
131	.604	-2.577	B
181	1.020	-2.584	B
151	.438	-2.651	A
48	.266	-2.696	B
65	1.024	-2.711	B
135	.425	-2.789	A
121	.743	-2.820	A
17	.716	-2.891	B
201	.310	-2.966	B

Table B
Test Reaction Items Used for the Perceived Test Difficulty and Test-Taking Anxiety and Motivation Scales, and Scoring Weights for Each Response

Scale and Item	Scoring Weight
Perceived Test Difficulty	
1. How often did you feel that the questions in the test were too easy for you?	
(1) Always	1
(2) Frequently	2
(3) Sometimes	3
(4) Seldom	4
(5) Never	5
6. In relation to your vocabulary ability, how difficult was the test for you?	
(1) Much too difficult	-1
(2) Somewhat too difficult	-2
(3) Just about right	-3
(4) Somewhat too easy	-4
(5) Much too easy	-5
9. How often were you sure that your answers to the questions were correct?	
(1) Almost always	1
(2) More than half of the time	2
(3) About half of the time	3
(4) Less than half of the time	4
(5) Almost never	5
13. How often did you feel that the questions in the test were too hard for you?	
(1) Always	-1
(2) Frequently	-2
(3) Sometimes	-3
(4) Seldom	-4
(5) Never	-5
Test-Taking Anxiety	
2. How did you feel while taking the test?	
(1) Very tense	-1
(2) Somewhat tense	-2
(3) Neither tense nor relaxed	-3
(4) Somewhat relaxed	-4
(5) Very relaxed	-5
Test-Taking Anxiety, *cont'd.*	
4. During testing, did you worry about how well you would do?	
(1) Not at all	1
(2) Very little	2
(3) Somewhat	3
(4) Fairly much so	4
(5) Very much so	5
7. Did nervousness while taking the test prevent you from doing your best?	
(1) Definitely	-1
(2) Probably	-2
(3) Not sure	-3
(4) Probably not	-4
(5) Definitely not	-5
11. Were you nervous while taking the test?	
(1) Not at all	1
(2) Somewhat	2
(3) Very little	3
(4) Moderately so	4
(5) Very much so	5
Test-Taking Motivation	
3. Did you feel challenged to do as well as you could on the test?	
(1) Not at all	1
(2) Very little	2
(3) Somewhat	3
(4) Fairly much so	4
(5) Very much so	5
10. Did you care how well you did on the test?	
(1) I cared a lot	-1
(2) I cared some	-2
(3) I cared a little	-3
(4) I cared very little	-4
(5) I didn't care at all	-5
12. Do you think that you could have done better on the test if you had tried harder?	
(1) I definitely could have	1
(2) I probably could have	2
(3) I'm not sure	3
(4) I probably couldn't have	4
(5) I definitely couldn't have	5

Table C
Ability Estimate ($\hat{\theta}$), Total Proportion Correct (T), and Proportion Correct for Each Student on Each of the Substrata (A, B) of the Nine-Stratum Test

			Stratum and Substratum																	
			1		2		3		4		5		6		7		8		9	
Student	$\hat{\theta}$	T	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
1	-1.01	.47	.86	.88	.77	.76	.60	.60	.49	.50	.41	.42	.34	.35	.27	.27	.25	.25	.26	.26
2	-1.87	.35	.70	.73	.51	.48	.39	.39	.34	.34	.29	.30	.27	.27	.23	.23	.23	.23	.23	.23
3	-.11	.61	.94	.95	.92	.92	.82	.81	.71	.71	.60	.62	.50	.52	.37	.37	.31	.31	.31	.31
4	-1.50	.40	.78	.81	.63	.60	.47	.47	.39	.40	.33	.34	.29	.30	.24	.24	.24	.23	.24	.24
5	-.28	.59	.93	.94	.90	.90	.78	.78	.67	.67	.56	.58	.46	.48	.34	.34	.30	.30	.30	.30
6	-.75	.51	.89	.91	.83	.82	.67	.67	.55	.56	.46	.47	.38	.39	.28	.29	.27	.27	.27	.27
7	-1.50	.40	.78	.81	.63	.60	.47	.47	.39	.40	.33	.34	.29	.30	.24	.24	.24	.24	.24	.24
8	-1.70	.38	.74	.77	.56	.53	.43	.42	.36	.36	.31	.31	.28	.28	.23	.23	.23	.23	.24	.23
9	1.45	.84	.98	.98	.98	.99	.96	.96	.93	.93	.89	.90	.84	.85	.79	.80	.65	.62	.50	.53
10	-.68	.52	.90	.91	.85	.83	.69	.69	.57	.58	.47	.49	.39	.40	.29	.29	.27	.27	.27	.27
11	-.72	.52	.90	.91	.84	.82	.68	.67	.56	.57	.46	.48	.38	.39	.29	.29	.27	.27	.27	.27
12	-.95	.48	.87	.89	.79	.77	.61	.61	.51	.51	.41	.43	.35	.36	.27	.27	.26	.26	.26	.26
13	.44	.70	.96	.97	.96	.96	.89	.89	.81	.82	.73	.74	.64	.66	.50	.52	.38	.38	.36	.37
14	-.94	.48	.87	.89	.79	.77	.62	.61	.51	.52	.52	.42	.43	.35	.36	.27	.27	.26	.26	.26
15	.41	.69	.96	.97	.95	.95	.89	.89	.81	.81	.72	.74	.63	.65	.50	.51	.37	.37	.36	.37
16	-.71	.52	.90	.91	.84	.83	.68	.68	.56	.57	.46	.48	.38	.40	.29	.29	.27	.27	.27	.27
17	-.89	.49	.88	.89	.80	.78	.63	.63	.52	.53	.43	.44	.36	.37	.27	.27	.26	.26	.26	.26
18	-.50	.55	.92	.93	.87	.87	.73	.73	.61	.62	.51	.53	.42	.44	.31	.31	.28	.28	.29	.28
19	.87	.76	.97	.98	.97	.97	.93	.93	.88	.88	.81	.82	.74	.76	.64	.65	.47	.46	.42	.43
20	-.11	.61	.94	.95	.92	.92	.81	.81	.70	.71	.60	.62	.50	.52	.37	.37	.31	.31	.31	.31
21	-1.63	.38	.75	.78	.58	.56	.44	.44	.37	.37	.31	.32	.28	.29	.24	.24	.23	.23	.24	.24
22	-.98	.48	.87	.88	.78	.76	.61	.60	.50	.51	.41	.43	.34	.35	.27	.27	.25	.25	.26	.26
23	-.43	.56	.92	.93	.88	.88	.75	.75	.63	.64	.53	.54	.43	.45	.32	.32	.28	.29	.29	.29
24	.93	.77	.97	.98	.97	.98	.93	.93	.88	.89	.82	.83	.75	.77	.66	.67	.49	.47	.42	.44
25	-.71	.52	.90	.91	.84	.83	.68	.68	.56	.57	.46	.48	.38	.40	.29	.29	.27	.27	.27	.27
26	1.00	.78	.98	.98	.97	.98	.94	.94	.89	.89	.83	.84	.77	.78	.68	.69	.51	.49	.43	.45
27	.71	.74	.97	.97	.97	.97	.92	.92	.85	.86	.78	.79	.70	.72	.59	.60	.43	.43	.40	.41
28	.09	.64	.95	.96	.94	.93	.85	.85	.75	.75	.65	.67	.55	.57	.41	.42	.33	.33	.33	.33
29	-.70	.52	.90	.91	.84	.83	.68	.68	.56	.57	.47	.48	.38	.40	.29	.29	.27	.27	.27	.27
30	-3.03	.26	.43	.46	.29	.27	.26	.26	.24	.24	.23	.23	.23	.22	.21	.21	.21	.21	.21	.21
31	-.02	.63	.95	.95	.93	.92	.83	.83	.73	.73	.62	.64	.52	.55	.39	.39	.32	.32	.32	.32
32	-1.60	.39	.76	.79	.59	.57	.45	.44	.38	.38	.32	.33	.29	.29	.24	.24	.23	.23	.24	.24
33	-.47	.56	.92	.93	.88	.87	.74	.74	.62	.63	.52	.54	.42	.44	.31	.32	.28	.28	.29	.29
34	.13	.65	.95	.96	.94	.94	.85	.85	.75	.76	.66	.67	.55	.58	.42	.43	.33	.34	.33	.34
35	-1.08	.46	.85	.87	.75	.74	.58	.58	.48	.48	.49	.41	.33	.34	.26	.26	.25	.25	.26	.25
36	-1.04	.47	.86	.88	.76	.75	.59	.59	.49	.49	.40	.41	.34	.35	.26	.26	.25	.25	.26	.26
37	-1.14	.45	.85	.86	.74	.72	.56	.56	.47	.47	.38	.40	.33	.33	.26	.26	.25	.25	.25	.25
38	.61	.72	.97	.97	.96	.96	.91	.91	.84	.85	.76	.78	.68	.70	.56	.57	.41	.41	.38	.39
39	-.24	.59	.93	.94	.91	.90	.79	.79	.68	.68	.57	.59	.47	.49	.35	.35	.30	.30	.30	.30
40	.47	.70	.96	.97	.96	.96	.90	.90	.82	.82	.73	.75	.64	.67	.51	.53	.38	.38	.37	.37
41	1.82	.88	.99	.99	.99	.99	.97	.97	.95	.95	.93	.93	.88	.89	.85	.86	.75	.73	.55	.60
42	-.89	.49	.88	.89	.80	.79	.63	.63	.52	.53	.43	.44	.36	.37	.27	.27	.26	.26	.26	.26
43	1.01	.78	.98	.98	.97	.98	.94	.94	.89	.90	.84	.84	.77	.78	.68	.70	.51	.49	.43	.45
44	-.11	.61	.94	.95	.92	.91	.81	.81	.70	.71	.60	.62	.50	.52	.37	.37	.31	.31	.31	.31
45	-.70	.52	.90	.91	.84	.83	.68	.68	.57	.57	.47	.48	.38	.40	.29	.29	.27	.27	.27	.27
46	-.28	.58	.93	.94	.90	.90	.78	.78	.67	.67	.56	.58	.46	.48	.34	.34	.29	.30	.30	.30
47	-.38	.57	.93	.93	.89	.88	.76	.76	.64	.65	.54	.56	.44	.46	.33	.33	.29	.29	.29	.29
48	1.29	.82	.98	.98	.98	.98	.95	.95	.92	.92	.87	.88	.81	.83	.75	.77	.60	.57	.47	.50
49	-2.34	.30	.58	.62	.38	.36	.32	.32	.28	.28	.25	.26	.25	.25	.22	.22	.22	.22	.22	.22
50	-.55	.54	.91	.92	.87	.86	.72	.72	.60	.61	.50	.52	.41	.43	.30	.31	.28	.28	.28	.28
51	.74	.74	.97	.97	.97	.97	.92	.92	.86	.86	.79	.80	.71	.73	.60	.61	.44	.43	.40	.41
52	.42	.69	.96	.97	.96	.95	.89	.89	.81	.81	.72	.74	.63	.65	.50	.51	.37	.38	.36	.37
53	-2.43	.30	.56	.60	.37	.35	.31	.31	.28	.28	.25	.25	.24	.24	.22	.22	.22	.22	.22	.22
54	-.91	.49	.88	.89	.80	.78	.63	.62	.52	.52	.42	.44	.35	.37	.27	.27	.26	.26	.26	.26
55	-1.23	.44	.83	.85	.71	.69	.54	.54	.45	.45	.37	.38	.32	.32	.25	.25	.25	.24	.25	.25
56	-1.73	.37	.73	.76	.55	.53	.42	.42	.36	.36	.30	.31	.28	.28	.23	.23	.23	.23	.24	.23
57	.27	.67	.96	.96	.95	.95	.87	.87	.78	.79	.69	.71	.59	.62	.46	.47	.35	.36	.35	.35
58	-.63	.53	.91	.92	.85	.84	.70	.70	.58	.59	.48	.50	.39	.41	.30	.30	.27	.27	.28	.28
59	-2.21	.32	.61	.65	.41	.39	.34	.34	.29	.30	.25	.27	.25	.25	.22	.22	.22	.22	.23	.22
60	-.57	.54	.91	.92	.86	.85	.71	.71	.60	.60	.49	.51	.40	.42	.30	.30	.28	.28	.28	.28
61	.22	.66	.96	.96	.94	.94	.87	.87	.77	.78	.68	.70	.58	.61	.44	.45	.35	.35	.34	.35
62	-.57	.54	.91	.92	.86	.85	.72	.71	.60	.60	.50	.51	.40	.42	.30	.30	.28	.28	.28	.28
63	-1.36	.42	.81	.83	.67	.65	.51	.50	.42	.42	.35	.36	.31	.31	.25	.25	.24	.24	.25	.24
64	-.23	.59	.93	.94	.91	.90	.79	.79	.68	.68	.57	.59	.47	.49	.35	.35	.30	.30	.30	.30
65	-1.68	.38	.74	.77	.57	.54	.43	.43	.36	.37	.31	.32	.28	.28	.23	.23	.23	.23	.24	.23
66	-.46	.56	.92	.93	.88	.87	.74	.74	.62	.63	.52	.54	.42	.44	.32	.32	.28	.28	.29	.29
67	-.08	.62	.94	.95	.92	.92	.82	.82	.71	.72	.61	.63	.50	.53	.37	.38	.31	.32	.32	.32
68	.54	.71	.97	.97	.96	.96	.90	.90	.83	.83	.75	.76	.66	.68	.53	.55	.39	.39	.38	.38
69	-.99	.48	.87	.88	.78	.76	.60	.60	.50	.50	.41	.42	.34	.35	.27	.27	.25	.25	.26	.26
70	-.06	.62	.94	.95	.92	.92	.82	.82	.72	.72	.61	.63	.51	.54	.38	.38	.31	.32	.32	.32
71	-.64	.53	.90	.92	.85	.84	.70	.69	.58	.59	.48	.50	.39	.41	.29	.30	.27	.27	.28	.28
72	.17	.65	.95	.96	.94	.94	.86	.86	.76	.77	.67	.68	.56	.59	.43	.44	.34	.34	.34	.34
73	.78	.75	.97	.97	.97	.97	.92	.92	.86	.87	.80	.81	.72	.74	.61	.63	.45	.44	.40	.42
74	.53	.71	.97	.97	.96	.96	.90	.90	.83	.83	.75	.76	.66	.68	.53	.55	.39	.39	.37	.38
75	-2.38	.30	.57	.61	.37	.35	.32	.31	.28	.28	.25	.25	.25	.24	.22	.22	.22	.22	.22	.22

(continued)

Table C *(continued)*

Ability Estimate ($\hat{\theta}$), Total Proportion Correct (T) and Proportion Correct for Each Student on Each of the Substrata (A, B) of the Nine-Stratum Test

			Stratum and Substratum																	
			1		2		3		4		5		6		7		8		9	
Student	$\hat{\theta}$	T	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
76	-.54	.54	.91	.92	.87	.86	.72	.72	.60	.61	.50	.52	.41	.43	.30	.31	.28	.28	.28	.28
77	-.50	.55	.92	.93	.87	.86	.73	.73	.61	.62	.51	.53	.42	.44	.31	.31	.28	.28	.29	.28
78	-.89	.49	.88	.89	.80	.78	.63	.63	.52	.53	.43	.44	.36	.37	.27	.27	.26	.26	.26	.26
79	-1.56	.39	.77	.80	.61	.58	.46	.46	.38	.39	.32	.33	.29	.29	.24	.24	.24	.23	.24	.24
80	-.97	.48	.87	.88	.78	.77	.61	.61	.50	.51	.41	.43	.35	.36	.27	.27	.26	.25	.26	.26
81	-3.08	.25	.42	.45	.28	.26	.26	.26	.24	.24	.23	.23	.23	.22	.21	.21	.21	.21	.21	.21
82	-.93	.49	.87	.89	.79	.78	.62	.62	.51	.52	.42	.44	.35	.36	.27	.27	.26	.26	.26	.26
83	.43	.70	.96	.97	.96	.96	.89	.89	.81	.82	.73	.74	.63	.66	.50	.52	.38	.38	.36	.37
84	.65	.73	.97	.97	.96	.97	.91	.91	.85	.85	.77	.78	.69	.71	.57	.59	.42	.41	.39	.40
85	-.60	.54	.91	.92	.86	.85	.71	.71	.59	.60	.49	.51	.40	.42	.30	.30	.27	.27	.28	.28
86	-.43	.56	.92	.93	.88	.88	.75	.75	.63	.64	.53	.54	.43	.45	.32	.32	.28	.29	.29	.29
87	-1.43	.41	.80	.82	.65	.63	.49	.49	.41	.41	.34	.35	.30	.30	.24	.24	.24	.24	.24	.24
88	-1.33	.43	.81	.83	.68	.66	.51	.51	.42	.43	.35	.36	.31	.31	.25	.25	.24	.24	.25	.24
89	-.87	.49	.88	.89	.81	.79	.64	.63	.53	.53	.43	.45	.36	.37	.27	.28	.26	.26	.26	.26
90	.39	.69	.96	.97	.95	.95	.89	.89	.81	.81	.72	.73	.62	.65	.49	.50	.37	.37	.36	.37
91	-.35	.57	.93	.94	.89	.89	.77	.76	.65	.66	.54	.56	.44	.47	.33	.33	.29	.29	.29	.29
92	-.03	.62	.95	.95	.93	.92	.83	.83	.72	.73	.62	.64	.51	.54	.38	.39	.32	.32	.32	.32
93	-.56	.69	.96	.97	.95	.95	.89	.89	.81	.81	.72	.73	.62	.65	.49	.50	.37	.37	.36	.37
94	-2.40	.30	.56	.60	.37	.35	.32	.31	.28	.28	.25	.25	.24	.24	.22	.22	.22	.22	.22	.22
95	-.88	.49	.88	.89	.80	.79	.64	.63	.52	.53	.43	.45	.36	.37	.27	.28	.26	.26	.26	.26
96	-.30	.58	.93	.94	.90	.89	.78	.78	.66	.67	.56	.57	.45	.48	.34	.34	.29	.30	.30	.30
97	-1.84	.36	.70	.74	.51	.49	.40	.40	.34	.34	.29	.30	.27	.27	.23	.23	.23	.23	.23	.23
98	-.64	.53	.90	.92	.85	.84	.70	.70	.58	.59	.48	.50	.39	.41	.30	.30	.27	.27	.28	.28
99	-1.26	.44	.83	.85	.70	.68	.53	.53	.44	.45	.36	.38	.32	.32	.25	.25	.24	.24	.25	.25
100	-.51	.55	.92	.93	.87	.86	.73	.73	.61	.62	.51	.53	.41	.43	.31	.31	.28	.28	.28	.28
101	.96	.77	.98	.98	.97	.98	.94	.94	.89	.89	.83	.84	.76	.77	.66	.68	.50	.48	.43	.45
102	-2.10	.33	.64	.68	.44	.42	.36	.35	.31	.31	.27	.27	.26	.26	.22	.22	.22	.22	.23	.22
103	1.36	.83	.98	.98	.98	.99	.96	.96	.93	.93	.88	.89	.83	.84	.77	.78	.63	.60	.48	.51
104	-.80	.51	.89	.90	.82	.81	.66	.65	.54	.55	.45	.46	.37	.38	.28	.28	.26	.26	.27	.27
105	-.75	.51	.89	.91	.83	.82	.67	.67	.55	.56	.46	.47	.38	.39	.29	.29	.27	.27	.27	.27
106	-.79	.51	.89	.90	.82	.81	.66	.66	.54	.55	.45	.46	.37	.38	.28	.28	.26	.26	.27	.27
107	-.23	.59	.93	.94	.91	.90	.79	.79	.68	.68	.57	.59	.47	.49	.35	.35	.30	.30	.30	.30
108	.27	.67	.96	.96	.95	.95	.87	.87	.78	.79	.69	.71	.59	.62	.46	.47	.35	.35	.35	.35
109	-.39	.57	.92	.93	.89	.88	.76	.76	.64	.65	.54	.55	.44	.46	.32	.33	.29	.29	.29	.29
110	-.38	.57	.93	.93	.89	.88	.76	.76	.64	.65	.54	.56	.44	.46	.33	.33	.29	.29	.29	.29
111	-.68	.52	.90	.91	.84	.83	.69	.68	.57	.58	.47	.49	.39	.40	.29	.29	.27	.27	.27	.27
112	-.27	.59	.93	.94	.90	.90	.78	.78	.67	.68	.56	.58	.46	.49	.34	.35	.30	.30	.30	.30
113	-.85	.50	.88	.90	.81	.79	.64	.64	.53	.54	.43	.45	.36	.37	.28	.28	.26	.26	.27	.26
114	1.06	.79	.98	.98	.98	.98	.94	.94	.90	.90	.84	.85	.78	.79	.69	.71	.53	.51	.44	.46
115	-.82	.50	.89	.90	.82	.80	.65	.65	.54	.54	.44	.46	.37	.38	.28	.28	.26	.26	.27	.27
116	.42	.69	.96	.97	.96	.95	.89	.89	.81	.81	.72	.74	.63	.65	.50	.51	.37	.38	.36	.37
117	.89	.76	.97	.98	.97	.97	.93	.93	.88	.88	.82	.82	.74	.76	.64	.66	.48	.46	.42	.43
118	-3.40	.24	.37	.39	.26	.24	.24	.24	.23	.23	.22	.22	.22	.22	.21	.21	.21	.21	.21	.21
119	-.88	.49	.88	.89	.80	.79	.64	.63	.52	.53	.43	.45	.36	.37	.27	.28	.26	.26	.26	.26
120	-.49	.55	.92	.93	.88	.87	.74	.73	.62	.62	.51	.53	.42	.44	.31	.31	.28	.28	.29	.29
121	.51	.71	.97	.97	.96	.96	.90	.90	.82	.83	.74	.76	.65	.68	.52	.54	.39	.39	.37	.38
122	2.07	.90	.99	.99	.99	.99	.98	.98	.96	.96	.94	.95	.90	.91	.89	.89	.81	.79	.60	.64
123	-1.95	.34	.68	.71	.48	.46	.38	.38	.32	.33	.28	.29	.27	.27	.23	.23	.23	.22	.23	.23
124	.75	.74	.97	.97	.97	.97	.92	.92	.86	.86	.79	.80	.71	.73	.60	.62	.44	.43	.40	.41
125	-2.08	.33	.65	.68	.45	.42	.36	.36	.31	.31	.27	.28	.26	.26	.22	.22	.22	.22	.23	.23
126	-.98	.48	.87	.88	.78	.76	.61	.60	.50	.51	.41	.43	.34	.35	.27	.27	.25	.25	.26	.26
127	.82	.75	.97	.97	.97	.97	.93	.93	.87	.87	.80	.81	.73	.75	.62	.64	.46	.45	.41	.42
128	-.14	.61	.94	.95	.92	.91	.81	.81	.70	.70	.59	.61	.49	.52	.36	.37	.31	.31	.31	.31
129	-1.71	.37	.74	.76	.56	.53	.42	.42	.36	.36	.30	.31	.28	.28	.23	.23	.23	.23	.24	.23
130	-.16	.60	.94	.95	.92	.91	.81	.81	.69	.70	.59	.61	.49	.51	.36	.37	.31	.31	.31	.31
131	.06	.64	.95	.96	.93	.93	.84	.84	.74	.75	.64	.66	.54	.57	.40	.41	.33	.33	.33	.33
132	-.57	.54	.91	.92	.86	.85	.72	.71	.60	.60	.49	.51	.40	.42	.30	.30	.28	.28	.28	.28
133	-1.32	.43	.81	.84	.68	.66	.51	.51	.43	.43	.35	.37	.31	.31	.25	.25	.24	.24	.25	.24
134	-2.08	.33	.65	.68	.45	.42	.36	.36	.31	.31	.27	.28	.26	.26	.22	.22	.22	.22	.23	.23
135	-1.01	.47	.86	.88	.77	.75	.60	.59	.49	.50	.40	.42	.34	.35	.27	.27	.25	.25	.26	.26
136	1.87	.88	.99	.99	.99	.99	.97	.97	.96	.96	.93	.93	.89	.89	.86	.87	.77	.74	.56	.61
137	-.73	.52	.90	.91	.84	.82	.68	.67	.56	.57	.46	.48	.38	.39	.29	.29	.27	.27	.27	.27
138	.35	.68	.96	.96	.95	.95	.88	.88	.80	.80	.71	.72	.61	.64	.48	.49	.36	.37	.36	.36
139	.25	.67	.96	.96	.95	.94	.87	.87	.78	.78	.69	.70	.59	.61	.45	.46	.35	.35	.34	.35
140	-1.05	.47	.86	.87	.76	.74	.59	.58	.48	.49	.40	.41	.34	.35	.26	.26	.25	.25	.26	.25
141	-.74	.51	.90	.91	.83	.82	.67	.67	.56	.56	.46	.48	.38	.39	.29	.29	.27	.27	.27	.27
142	.75	.74	.97	.97	.97	.97	.92	.92	.86	.86	.79	.80	.71	.73	.60	.62	.44	.43	.40	.41
143	-.47	.56	.92	.93	.88	.87	.74	.74	.62	.63	.52	.53	.42	.44	.31	.32	.28	.28	.29	.29
144	-.96	.48	.87	.88	.79	.77	.61	.61	.51	.51	.41	.43	.35	.36	.27	.27	.26	.25	.26	.26
145	-1.00	.47	.86	.88	.77	.76	.60	.60	.49	.50	.41	.42	.34	.35	.27	.27	.25	.25	.26	.26
146	-1.01	.47	.86	.88	.77	.75	.60	.59	.49	.50	.40	.42	.34	.35	.27	.27	.25	.25	.26	.26
147	-.75	.51	.89	.91	.83	.82	.67	.67	.55	.56	.46	.47	.38	.39	.28	.29	.27	.27	.27	.27
148	-.14	.61	.94	.95	.92	.91	.81	.81	.70	.70	.59	.61	.49	.52	.36	.37	.31	.31	.31	.31
149	1.29	.82	.98	.98	.98	.98	.95	.95	.92	.92	.87	.88	.82	.83	.75	.77	.60	.57	.47	.50
150	-.04	.62	.94	.95	.93	.92	.83	.83	.72	.73	.62	.64	.51	.54	.38	.39	.32	.32	.32	.32
151	-.13	.61	.94	.95	.92	.91	.81	.81	.70	.71	.60	.62	.49	.52	.37	.37	.31	.31	.31	.31

DISTRIBUTION LIST

Navy

1 Dr. Ed Aiken
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 MR. MAURICE CALLAHAN
Pers 23a
Bureau of Naval Personnel
Washington, DC 20370

1 Dr. Richard Elster
Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93940

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Dr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 CAPT. D.M. GRAGG, MC, USN
HEAD, SECTION ON MEDICAL EDUCATION
UNIFORMED SERVICES UNIV. OF THE
HEALTH SCIENCES
6917 ARLINGTON ROAD
BETHESDA, MD 20014

1 CDR Robert S. Kennedy
Naval Aerospace Medical and
Research Lab
Box 29407
New Orleans, LA 70189

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

1 CHAIRMAN, LEADERSHIP & LAW DEPT.
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMYY
ANNAPOLIS, MD 21402

1 CAPT Richard L. Martin
USS Francis Marion (LPA-Z49)
FPO New York, NY 09501

1 Dr. James McBride
Code 301
Navy Personnel R&D Center
San Diego, CA 92152

1 Library
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 OFFICE OF CIVILIAN PERSONNEL
(CODE 26)
DEPT. OF THE NAVY
WASHINGTON, DC 20390

1 JOHN OLSEN
CHIEF OF NAVAL EDUCATION &
TRAINING SUPPORT
PENSACOLA, FL 32509

1 Psychologist
ONR Branch Office
495 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 200
Arlington, VA 22217

1 Code 436
Office of Naval Research
Arlington, VA 22217

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
OFFICE OF NAVAL RESEARCH BRANCH
223 OLD MARYLEBONE ROAD
LONDON, NW, 15TH ENGLAND

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco, CA 96503

1 Office of the Chief of Naval Operations
Research, Development, and Studies Branc
(OP-102)
Washington, DC 20350

1 Scientific Advisor to the Chief of
Naval Personnel (Pers-Or)
Naval Bureau of Personnel
Room 4410, Arlington Annex
Washington, DC 20370

1 LT Frank C. Petho, MSC, USNR (Ph.D)
Code L51
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 Roger W. Remington, Ph.D
Code L52
NAMRL
Pensacola, FL 32508

1 Mr. Arnold Rubenstein
Naval Personnel Support Technology
Naval Material Command (08T244)
Room 1044, Crystal Plaza #5
2221 Jefferson Davis Highway
Arlington, VA 20360

1 Dr. Worth Scanland
Chief of Naval Education and Training
Code N-5
NAS, Pensacola, FL 32508

1 A. A. SJOHOLM
TECH. SUPPORT, CODE 201
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr. Robert Smith
Office of Chief of Naval Operations
OP-987E
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 CDR Charles J. Theisen, JR. MSC, USN
Head Human Factors Engineering Div.
Naval Air Development Center
Warminster, PA 18974

1 W. Gary Thomson
Naval Ocean Systems Center
Code 7132
San Diego, CA 92152

1 Dr. Ronald Weitzman
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 HQ USAREUE & 7th Army
ODCSOPS
USAAREUE Director of GED
APO New York 09403

1 DR. RALPH DUSEK
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Michael Kaplan
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Beatrice J. Farr
Army Research Institute (PERI-OK)
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Milt Maier
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Harold F. O'Neil, Jr.
ATTN: PERI-OK
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Director, Training Development
U.S. Army Administration Center
ATTN: Dr. Sherrill
Ft. Benjamin Harrison, IN 46218

1 Dr. Frederick Steinheiser
U. S. Army Reserch Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/PED
Brooks AFB, TX 78235

1 Air University Library
AUL/LSE 76/443
Maxwell AFB, AL 36112

1 Dr. Philip De Leo
AFHRL/TT
Lowry AFB, CO 80230

1 DR. G. A. ECKSTRAND
AFHRL/AS
WRIGHT-PATTERSON AFB, OH 45433

1 Dr. Genevieve Haddad
Program Manager
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 CDR. MERCER
CNET LIAISON OFFICER
AFHRL/FLYING TRAINING DIV.
WILLIAMS AFB, AZ 85224

1 Dr. Ross L. Morgan (AFHRL/ASR)
Wright -Patterson AFB
Ohio 45433

1 Dr. Roger Pennell
AFHRL/TT
Lowry AFB, CO 80230

1 Personnel Analysis Division
HQ USAF/DPXXA
Washington, DC 20330

1 Research Branch
AFMPC/DPMYP
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/PED
Brooks AFB, TX 78235

1 Dr. Marty Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Jack A. Thorpe, Capt, USAF
Program Manager
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 Brian K. Waters, LCOL, USAF
Air University
Maxwell AFB
Montgomery, AL 36112

Marines

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 Mr. Richard Lanterman
PSYCHOLOGICAL RESEARCH (G-P-1/62)
U.S. COAST GUARD HQ
WASHINGTON, DC 20590

1 Dr. Thomas Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria, VA 22314
Attn: TC

1 Dr. Dexter Fletcher
ADVANCED RESEARCH PROJECTS AGENCY
1400 WILSON BLVD.
ARLINGTON, VA 22209

1 Dr. William Graham
Testing Directorate
MEPCOM
Ft. Sheridan, IL 60037

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 MAJOR Wayne Sellman, USAF
Office of the Assistant Secretary
of Defense (MRA&L)
3B930 The Pentagon
Washington, DC 20301

Civil Govt

1 Dr. Susan Chipman
Basic Skills Program
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. William Gorham, Director
Personnel R&D Center
Office of Personnel Managment
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph I. Lipson
Division of Science Education
Room W-638
National Science Foundation
Washington, DC 20550

1 Dr. John Mays
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Arthur Melmed
National Intitute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Dr. Lalitha P. Sanathanan
Environmental Impact Studies Division
Argonne National Laboratory
9700 S. Cass Avenue
Argonne, IL 60439

1 Dr. Jeffrey Schiller
National Institute of Education
1200 19th St. NW
Washington, DC 20208

1 Dr. Thomas G. Sticht
Basic Skills Program
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Managment
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Dr. Erling B. Anderson
University of Copenhagen
Studiestraedt
Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Alan Baddeley
Medical Research Council
Applied Psychology Unit
15 Chaucer Road
Cambridge CB2 2EF
ENGLAND

1 Dr. Isaac Dejar
Educational Testing Service
Princeton, NJ 08450

1 Dr. Warner Birice
Streitkraefteamt
Rosenberg 5300
Bonn, West Germany D-5300

1 Dr. R. Darrel Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Dr. Nicholas A. Bond
Dept. of Psychology
Sacramento State College
600 Jay Street
Sacramento, CA 95819

1 Dr. David G. Bowers
Institute for Social Research
University of Michigan
Ann Arbor, MI 48106

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. William Coffman
Iowa Testing Programs
University of Iowa
Iowa City, IA 52242

1 Dr. Allan M. Collins
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, Ma 02138

1 Dr. Meredith Crawford
Department of Engineering Administration
George Washington University
Suite 805
2101 L Street N. W.
Washington, DC 20037

1 Dr. Hans Cronbag
Education Research Center
University of Leyden
Boerhaavelaan 2
Leyden
The NETHERLANDS

1 MAJOR I. N. EVONIC
CANADIAN FORCES PERS. APPLIED RESEARCH
1107 AVENUE ROAD
TORONTO, ONTARIO, CANADA

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Gerhardt Fischer
Liebigasse 5
Vienna 1010
Austria

1 Dr. Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organ.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Ross Greene
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

1 Dr. Alan Gross
Center for Advanced Study in Education
City University of New York
New York, NY 10036

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
Department of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Carl J. Jensema
Gallaudet College
Kendall Green
Washington, DC 20002

1 Dr. Arnold F. Kanarick
Honeywell, Inc.
2600 Ridgeway Pkwy
Minneapolis, MN 55413

1 Dr. John A. Keats
University of Newcastle
Newcastle, New South Wales
AUSTRALIA

1 LCOL. C.R.J. LAFLEUR
PERSONNEL APPLIED RESEARCH
NATIONAL DEFENSE HQS
101 COLONEL BY DRIVE
OTTAWA, CANADA K1A OK2

1 Dr. Michael Levine
Department of Educational Psychology
University of Illinois
Champaign, IL 61820

1 Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat
Groningen
NETHERLANDS

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77025

1 Dr. Sam Mayo
Loyola University of Chicago
Chicago, IL 60601

1 Dr. Melvin R. Novick
Iowa Testing Programs
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analysis
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. STEVEN M. PINE
4950 Douglas Avenue
Golden Valley, MN 55416

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

1 MIN. RET. M. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 161
53 BONN 1, GERMANY

1 Dr. Peter B. Read
Social Science Research Council
605 Third Avenue
New York, NY 10016

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
12 Hill Hall
Columbia, MO 65201

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Donald Rubin
Educational Testing Service
Princeton, NJ 08450

1 Dr. Larry Rudner
Gallaudet College
Kendall Green
Washington, DC 20002

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 Dr. Kazao Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 982
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
Florida Technological University
Orlando, FL 32816

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN
THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Office of Data Analysis Research
Educational Testing Service
Princeton, NJ 08541

1 Dr. Kikumi Tatsuoka
Computer Based Education Research
Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. Maurice Tatsuoka
Department of Educational Psychology
University of Illinois
Champaign, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Dept. of Statistics
University of MIssouri
Columbia, MO 65201

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Bureau of Social SCience Research
1990 M Street, N. W.
Washington, DC 20036

1 Dr. Phyllis Weaver
Graduate School of Education
Harvard University
200 Larsen Hall, Appian Way
Cambridge, MA 02138

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Dr. Wolfgang Wildgrube
Streitkraefteamt
Rosenberg 5300
Bonn, West Germany D-5300

1 Dr. Robert Woud
School Examination Department
University of London
66-72 Gower Street
London WCIE 6EE
ENGLAND

1 Dr. Karl Zinn
Center for research on Learning
and Teaching
University of Michigan